Advances in Experimental Medicine and Biology 896

M. Cristina Vega *Editor*

Advanced Technologies for Protein Complex Production and Characterization

Springer

Advances in Experimental Medicine and Biology

Volume 896

More information about this series at http://www.springer.com/series/5584

M. Cristina Vega
Editor

Advanced Technologies for Protein Complex Production and Characterization

Springer

Editor
M. Cristina Vega
Center for Biological Research
Spanish National Research Council (CIB-CSIC)
Madrid, Spain

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-3-319-27214-6 ISBN 978-3-319-27216-0 (eBook)
DOI 10.1007/978-3-319-27216-0

Library of Congress Control Number: 2016939120

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Complex Biologics*: *Engineering the Tools and the Hosts. Proteins are at the core of virtually all structures and functions in living organisms: they are responsible for putting the genetic information into action. Proteins catalyze a myriad of chemical processes, assemble and maintain a seemingly unending repertoire of architectures—ranging from viral capsids to the entire cytoskeleton—generate movement and orchestrate cell division, mediate intracellular signaling cascades and intercellular communication and synchronization, and probably even catalyze our thoughts and store our memories. Moreover, many of our present-day therapeutic drugs are proteins and peptides (e.g., antibodies, hormones, antibiotics), and the role of protein drugs in combating disease is predicted to keep rising exponentially. In addition, a sizable portion of the biotechnology sector that delivers food and commodities and recycles waste will critically depend on the discovery and targeted engineering of novel proteins with customized enzymatic functionalities (e.g., glycosidases and synthetases, lipases, oxidoreductases, and others). On a larger industrial and societal scale, cleaner and sustainable "green" energy sources are being called upon to replace fossil fuels, nuclear power, and their long-term detrimental effects on ecology, populations, and human health.

The recent genomics and proteomics Big Data revolution has transformed our understanding of cellular function. Earlier, single catalytic units (proteins and other biomolecules) were the focus of attention—this has changed dramatically. It has become evident that the main actors of biological function in cells and organisms are intricate multicomponent interaction networks involving the cooperation of several to many functional units. This has profound implications for basic and applied research, directly impacting developments in biomedicine and the pharmaceutical and biotechnology industries. To understand biological function more fully and create better drugs and treatments for diseases will rely on new and powerful technologies, designed to recapitulate complex and multimodal biological functionalities in vitro and in vivo for the discovery of the underlying molecular structures and activities—ideally at atomic resolution, in meaningful physiological contexts. To address these challenges, significant resources have been invested in creating and deploying new technologies that can relieve the imposing bottlenecks. This present volume of *Advances in Experimental Medicine and Biology* provides a nonexhaustive selection of recent developments.

Multicomponent complexes that catalyze nearly all cellular and organismal functions comprise proteins, sugars, lipids, RNA, and/or DNA; they can be constitutively stable or assemble transiently at certain times and locations, upon particular stimuli. Efficient multicomponent protein and protein-nucleic acid co-production methods have been developed and validated to enable recombinant production of multicomponent cellular machines in a variety of heterologous host organisms. Simultaneous production of several to many polypeptide chains that gather to form (important parts of) a complex often is frequently found to improve yields and activities, often dramatically. Indeed, coproduction with partner molecules may at times be the only way to produce a particular protein catalyst for uses in research and development that require meaningful amounts of material—from a few milligrams in research laboratories to many grams in bioprocessing and for pharmacological applications. For many purposes, *Escherichia coli* is a well-suited heterologous host, and an overwhelming majority of the protein specimens that have been produced recombinantly exploited this versatile organism. Substantial engineering has been devoted to improve *E. coli* production and provide additional and useful functionalities. More recently, eukaryotic platforms for heterologous production have become increasingly utilized because of the advantages that these systems provide for producing eukaryotic, notably human proteins of interest and their complexes. These advantages include authentic processing and targeting, post-translational modifications, specific chaperone capabilities, eukaryotic membranes structures and compositions, among others. Furthermore, eukaryotic platforms are becoming more amenable and technically less demanding due to the streamlining of protocols and the availability of new reagents generated by state-of-the-art synthetic biology techniques and approaches. Expression systems that are particularly tailored for efficiently producing multicomponent assemblies have been devised in mammalian and insect cells with native-like activities. Notably, the MultiBac baculovirus/insect cell expression system has rendered accessible a large number of hitherto elusive biologically and pharmaceutically valuable protein complex specimens. Fungal hosts such as *Saccharomyces cerevisiae* and *Pichia pastoris* have also proved to offer cost-effective and versatile solutions for complex protein production and the engineering of metabolic pathways, with significant efforts underway to increase their usefulness and enhance their yields.

The present volume is the result of a combined initiative by Springer Press and the ComplexINC collaborative project. ComplexINC is a European Commission Framework Program 7 high-tech consortium combining leading expertise for developing new technologies and production tools for complex protein biologics. ComplexINC is moreover dedicated to the wide dissemination of these next-generation technologies in order to accelerate academic and industrial research and development in the European Research Area (ERA). This volume of *Advances in Experimental Medicine and Biology* is organized in seven parts. The introductory Part I (Chaps. 1 and 2) sets the context for protein complex production. Parts II–V cover fundamentals and describe practical approaches pertaining to selected prokaryotic and eukaryotic heterologous expression hosts. These sections also highlight technologies that offer unique opportunities, such as cell-free systems or plant-based systems as organismal production factories (Chaps. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18). Finally, Parts VI and VII review in vitro reconstitution techniques and a selection of powerful biophysical techniques for the characterization of multicomponent biological machines, including X-ray scattering (SAXS) and nuclear magnetic resonance (NMR) (Chaps. 19, 20, 21, and 22).

Proteins, their complexes, and their activities are central to modern biology, biotechnology, and biomedicine, and their heterologous production is often a vital prerequisite for discovering their function in health and disease. We strove to compile contributions that are both exciting and accessible; and we are hopeful to meet the interest of a large and diverse audience, from university students to junior and senior scientists, in academia and industry, in fundamental and applied research alike. We anticipate that the wide and successful implementation of new tools and advanced technologies to produce novel complex protein biologics, some of which are presented here, will critically contribute to new biomedical insights, next-generation drugs and, ultimately, to new and better therapeutic intervention strategies, for the benefit of all.

Madrid, Spain M. Cristina Vega

Grenoble, France Francisco J. Fernández
July 2015 Imre Berger

Contents

Part I Introduction

Part II Prokaryotic Expression Hosts

Part III Lower Eukaryotic Expression Hosts

Part IV Higher Eukaryotic Expression Hosts

Part V Plant Expression

Part VI Complex Reconstitution

Part VII Biophysical Methods to Assess Protein Complexes

About the Editor

M. Cristina Vega, PhD, is a Group Leader at the Chemical and Physical Biology Department of the Center for Biological Research (CIB-CSIC) in Madrid, Spain. Dr Vega's lab applies biochemical, biophysical and macromolecular X-ray diffraction methods to investigate how pathogens evade the immune system defence mechanisms. In addition, she designs eukaryotic, yeast-based expression platforms tailored to protein/protein complexes with a focus on cost efficiency and improved performance.

Part I

Introduction

Protein Complex Production from the Drug Discovery Standpoint

Ismail Moarefi

Abstract

Small molecule drug discovery critically depends on the availability of meaningful *in vitro* assays to guide medicinal chemistry programs that are aimed at optimizing drug potency and selectivity. As it becomes increasingly evident, most disease relevant drug targets do not act as a single protein. In the body, they are instead generally found in complex with protein cofactors that are highly relevant for their correct function and regulation. This review highlights selected examples of the increasing trend to use biologically relevant protein complexes for rational drug discovery to reduce costly late phase attritions due to lack of efficacy or toxicity.

Keywords

Protein complexes • Structure based drug discovery • Small molecule drug discovery • Proteomics • Protein complex expression

1.1 Introduction

The traditional approach to target based small molecule drug discovery strongly depends on the *in vitro* screening of disease relevant proteins as single polypeptide chains. In many cases even only the catalytic domain of a target protein is used for hit finding and optimization. However, this reductionist approach has its inherent limitations. *In vivo*, most of the disease targets being pursued by the pharmaceutical industry are not proteins that act in isolation. They are rather acting as part of multi-protein complexes. Within these complexes, activities are strongly modulated by interacting regulatory subunits. The interrogation of isolated domains and single polypeptide chains *in vitro* thus generally only poorly reflects the disease relevant state of the same protein in the human body. It is therefore no surprise that many drug candidates discovered in this way turn out to be costly failures during clinical trials, either due to lack of efficacy or adverse side effects. With the emergence of powerful

I. Moarefi (✉)
Crelux GmbH, Am Klopferspitz 19a,
D82152 Martinsried, Germany
e-mail: moarefi@crelux.com

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_1

proteomic analysis tools we are increasingly refining our structural and functional understanding of multi-protein complexes. The higher level of insight into the relevance of studying disease relevant proteins in their native multi-protein complex context is increasingly being applied to *in vitro* screening and structure-based drug discovery. This shift has the potential to lead the drug discovery industry to enhanced productivity in early phase small-molecule drug discovery and to higher success rates of investigational new drugs during clinical trials.

1.2 Some Key Requirements for the Production of Protein Complexes for Drug Discovery

Most cellular processes are based on the tightly regulated action of protein machines that generally are assembled as multi-protein complexes [1]. Recent proteomic analyses have convincingly shown that a large number of multi-protein complexes that are currently being identified by mass spectrometry and other powerful proteomics tools are completely un-annotated. While large multiprotein complexes have been relatively well characterized, a surprising number of complexes that are made of by five or fewer proteins appear to be completely undocumented and have never been described in the literature. These findings are somewhat surprising, given the large prevalence of well characterized human disease target proteins in exactly these smaller protein complexes [2]. What this means is that we are having only a very incomplete picture of the structure and regulated function of most of our disease targets at best. The apparent lack of vital information clearly highlights the need to thoroughly identify the relevant interaction partners of disease targets by state-of-the-art proteomics [1, 3] in cases where there is a lack of available information on the disease target of interest. Determination of the subunit composition, stoichiometry and status of posttranslational modifications ideally are the starting point for cloning, expression and purification of the disease target as a relevant protein complex in the desired functional state. This opens the way towards a detailed study of structure, function and regulation *in vitro* as the prerequisite to finding and optimizing small molecule modulators that are not only functional *in vitro*, but more importantly, also active *in vivo*.

Many choices of expression systems for multiprotein complexes are nowadays available and the most suitable one can be selected to fit the particular needs of a small molecule drug discovery campaign project. Important factors are, among others: speed, reliability, reproducibility, scalability and cost-of-goods.

1.2.1 Speed

How long does it take to arrive from the concept phase at successful complex production? Time is an extremely important factor in every drug discovery campaign being pursued in industry. Generally, the faster the protein complex can be cloned, expressed, purified and tested *in vitro*, the better. The activity of the overexpressed and purified protein complexes needs to be thoroughly characterized as early as possible. Importantly, *in vitro* assays need to be established and validated before a screening campaign can be initiated. Frequently the originally conceived expression constructs need to be further refined and optimized after the initial results are obtained to tune parameters like expression level and *in vitro* activity. Fine-tuning generally requires additional rounds of construct optimization and hence additional time required to obtain the desired protein complex in the correct activity state. Therefore the expression system of choice in the ideal case is amenable to automation and parallelization. This enables the researcher to test a large number of constructs and combinations in the first round. That way time-consuming iteration steps can be reduced to a minimum, albeit at the price of having to design and test a large number of construct hypotheses at an early stage.

1.2.2 Reliability and Reproducibility

Any expression system for protein complexes has to be particularly stable with respect to final yield, subunit composition and frequently the presence of the correct post-translational modifications in order to be dependable. Expression and downstream processing protocols should be stable enough to reliably deliver high quality protein complexes and maximized yields. Batch-to-batch variations have to be as low as possible in order to ensure dependable and reproducible results in biophysical and biochemical assays. For structure-based drug discovery activities, the purified complexes should crystallize reliably. Batch-to-batch variations should be minimal and not influence co-crystallization outcomes such as ability to grow crystals, presence or absence of small molecule ligands in the structures or resolution limits.

1.2.3 Scalability

During the drug discovery process a broad range of biochemical and biophysical assays are employed for hit-finding and hit-validation as well as hit- and lead-optimization. Depending on the assay format and number of data points to be measured, protein requirements for assays can readily exceed the 100 mg range [4]. In large high-throughput screening campaigns small molecule libraries of 2–7 million small-molecule compounds are routinely tested *in vitro* using biochemical assays. While each data point usually only requires minute amounts of purified protein complex, the large number of data points that need to be recorded results in an overall need of pure and active protein complex in the range of 10–100 mg. The result of such massive screening campaigns generally is a large number of initial hits that subsequently need to validated and confirmed by a range of biochemical and biophysical assays. If NMR based hit finding and validation methods are chosen, each individual measurement requires large amounts of protein that may seriously limit the number of experiments that can be carried out due to protein complex availability constraints. Clearly, not only the initial hit finding campaigns can pose considerable challenges to high quality protein supply. Biophysical methods that yield high quality data, such as ITC (isothermal titration calorimetry) and DSC (differential scanning calorimetry) are increasingly used for hit validation and thorough compound characterization during hit-to-lead programs to help the medicinal chemist in selecting and optimizing the most promising small molecules. These methods require large amounts of protein per data point, usually in the milligram range. Taken together, hit-finding, followed by hit-validation and hit-to-lead programs, which are the basic steps of lead discovery programs, can readily require high quality protein complexes in gram amounts during their execution [4].

1.2.4 Cost-of-Goods

Given the described requirements, which are to be met by the ideal expression system for protein complexes to efficiently power drug discovery programs, it becomes apparent that cost of goods is an important yet not the all important factor in the ultimate choice. Traditionally *E. coli* did serve as an economical workhorse for the production of single chain target proteins for *in vitro* testing whenever possible. However, protein complexes are posing additional challenges to an expression system in that generally large proteins have to be synthesized, correctly folded and correctly assembled with the right stoichiometry and the correct post-translational modifications being present in the final product. The move to the increasing use of protein complexes rather than single chains, therefore means that alternative expression systems, such as insect cell or mammalian cell based systems are given preference over their prokaryotic counterparts. The main reasons are that eukaryotic cells can provide more suitable expression and folding pathways for eukaryotic proteins that may require molecular chaperones and other factors responsible for correct folding and correct complex assembly *in vivo*.

Taking the above parameters into account, an expression system can usually be designed, implemented and tested that delivers the target protein of interest in a complex with the correct subunit composition and stoichiometry to ensure native-like activity and regulation. After purification, assay set-up and extensive *in vitro* characterization, the recombinant protein complex is ready for a range of biochemical and biophysical *in vitro* screens for hit finding and optimization. For structure-based hit optimization, complex structures can be obtained using crystallization optimized protein constructs.

In the following, a number of examples will be given in which protein complexes have been used for target based drug discovery. In some of the cases, the target protein of interest can only be generated with appropriate subunits being co-expressed. In other instances, the generated protein complexes have very different properties when compared to the single subunit target protein that contains the catalytic domain.

1.3 Cyclin Dependent Protein Kinases

There are about 520 distinct protein kinases in the human proteome [5]. Many of these kinases and their kinase domain regions are catalytically competent after auto-phosphorylation. This auto-phosphorylation typically that takes place in the presence of ATP, when the single protein chains are expressed, purified and tested in isolation. Other kinases depend on activation by phosphorylation *in trans* during transient interaction with upstream activating kinases and again can be catalytically active as single chains. Cyclin-dependent kinases (CDKs) are protein kinases that depend on the stable binding of specific cyclins to switch from an inactive conformation into a CDK/cyclin complex with basal kinase activity that in turn can be fully activated by phosphorylation of the CDK activation segment by CAK (Cdk Activating Kinase) that itself is a CDK7/cyclinH complex (reviewed in [6]).

Deregulated CDK activation is a common cause for a number of proliferative diseases, making many of these aberrantly activated CDKs very attractive targets for the development of selective ATP competitive inhibitors targeting the activated form. CDK2 was among the first kinases to be produced as a stable complex with cyclins. The purified enzyme can be obtained in the active form and is being used to screen for small molecule inhibitors *in vitro* [7]. For structure-based drug discovery, CDK2 can be reliably crystallized as a single, inactive, protein both as apo-protein as well as in complex with small-molecule ligands, mostly inhibitors that recognize and target the inactive conformation [6]. For small molecule inhibitors that bind and inhibit the activated form of CDK2, the most frequently used system for co-crystallization makes use of CDK2 in complex with a cyclinA fragment that contains the regions required for stable CDK interaction. For the first structures of CDK2 un-phosphorylated in complex with cyclinA, the two proteins were produced as separate chains in different hosts (cyclinA in *E. coli* and CDK2 in insect cells) and the complex formed using the two purified proteins prior to crystallization [8]. For the first CDK2/cyclinA complex in the fully active conformation the same expression strategy was used and the complex phosphorylated *in vitro*, during protein purification, by purified CAK (CDK7/cyclinH). CAK in turn can be generated by co-expression of CDK7/cyclinH in insect cells [9]. To date there are more that 140 entries in the PDB that contain human CDK2. More than 50 of these structures are CDK2/cyclinA/small molecule ternary complexes. This highlights the robustness and relevance of these systems for structure-based drug discovery. In the meantime, similar expression systems have been developed for disease relevant CDK/cyclin complexes and are being used for *in vitro* assays and for obtaining small molecule complex structures. Among them are CDK4/cyclinD1 [10], CDK6/Vcyclin (a viral cyclin from herpes virus) [11], CDK9/cyclinT1 [12] and CDK12/cyclinK [13].

1.4 Adenosine 5′-Monophosphate Activated Protein Kinases (AMPKs)

AMPKs are the master sensors that closely monitor the cellular energy status and metabolic stress and rapidly trigger cellular responses via the modulation of the phosphorylation state of their downstream target proteins (reviewed in [14]). They are playing a key role in regulating the whole body energy homeostasis and are attractive drug targets in a range of metabolic diseases, such as diabetes, neurodegenerative diseases and cancer [15]. AMPK basically works as an energy sensor by being able to monitor and measure the cellular concentrations of ATP, ADP and AMP. AMPKs are hetero-trimeric Serine/Threonine Protein kinases that are assembled from 1 alpha, 1 beta and 1 gamma subunit. There are 2 different types of α, 2 different types of β and 3 different types of γ subunits in the human proteome. This means that by combining the subunits in all possible combinations a total of 12 different AMPK isoforms can be assembled, each having unique properties. The canonical kinase domain is located in the N-terminal region of the α-subunit. The kinase activity is significantly enhanced upon phosphorylation of a threonine residue in the activation segment of the kinase domain. Upstream activating kinases that directly phosphorylate that threonine are liver kinase B1 (LKB1), calcium/calmodulin-dependent protein kinase kinase b (CaMKKb) and mammalian transforming growth factor β activated kinase 1 (TAK1) among others. The kinase domain of the AMPK α-subunit in isolation is completely inactive *in vitro*. The crystal structure of the isolated kinase domain of the α2 subunit has been solved (PDB code 2HD6) and shown to be in a catalytically inactive conformation. Fully regulated AMPK can only be generated when the hetero-trimeric complex is expressed and assembled. AMPK function critically depends on the concerted interaction of all three subunits as a function of cellular ATP levels and on activation by its upstream kinase. Phosphorylation of the AMPK activation segment is at least in part regulated by modulation of the accessibility of the threonine, which is in turn regulated by the γ-subunit that is able to detect, differentiate and quantify AMP, ADP and ATP upon binding to it. In order to obtain functional and correctly regulated AMPK isoforms for drug discovery, expression systems had to be established that yield soluble AMPK protein with the correct and stoichiometric subunit composition. The most successful and commonly used strategy has turned out to rely on expression of all three subunits from a tri-cistronic expression vector in *E. coli* cells [16]. For many AMPK isoforms, the final protein complex purified from this system could be demonstrated to be monodisperse and correctly assembled into a 1:1:1 trimer with the desired subunit composition. Determination of the activation state of the α-subunit shows that the threonine in the activation segment is not phosphorylated, in line with the absence of the activating kinases from *E. coli*. However, in some cases phosphorylation at different sites can be detected for example on the γ3 subunit when α2β3γ3 are expressed in *E. coli* [17]. Non-activated AMPK expressed and purified from *E. coli* can generally be activated *in vitro* using purified upstream kinases such as CaMKKb and LKB1. The activated AMPK complexes are successfully used for biochemical *in vitro* assays to identify and optimize activators or inhibitors [15]. For structure-based drug discovery, crystallization optimized constructs are being used where flexible regions are removed, which otherwise prevent the protein complex from forming well ordered and sufficiently diffracting crystals. A number of different AMPK isoforms have been successfully crystallized to date and their structures contributed significantly to expand our understanding of the intricacies of AMPK regulation [18] and activation by small molecule activators [9, 19] for this very promising family of drug targets.

1.5 Phosphoinositide-3 Kinases (PI3 Kinases)

PI3 Kinases (PI3Ks) are another class of important human drug targets. They are key elements of cellular signaling processes that control cellular survival, growth and proliferation. Their 3-kinase activity converts phosphatidylinositol (4,5) bisphosphate (PIP2) into the active phosphatidylinositol (3,4,5) trisphosphate (PIP3) at the plasma membrane [10, 20]. This signal is then transmitted and amplified via downstream enzyme cascades that include, among others, Serine/Threonine kinases such AKT (also known as PKB). Upregulation of PI3 kinases contributes to the development of solid tumors in a broad range of diseases such as breast, ovarian, colon and gastric cancers [21]. Type I PI3 kinases are hetero-dimers. The PI kinase activities of type IA PIK isoforms reside in the a, b and d, p110 chains. These 110 kDa chains are each interacting with a regulatory 85-kDa protein (p85a). The interaction between p110 and p85a inhibits the catalytic activity of the p110 subunit in the absence of upstream activating signals. While the catalytic domain of the type 1B p110g chain can be readily expressed and crystallized in the presence of small-molecule inhibitors, soluble type IA p110 protein expression requires co-expression with p85a. Coexpression of full length p110a with full-length p85a was successfully used to generate a complex suitable for crystallography [22]. Soluble, recombinant expression seems to minimally require the "inter"-SH2 (iSH2) domain of p85a. A construct where the p85a nSH2-iSH2 domains are fused to the N-terminus of the full length p110a catalytic domain is a constitutively active enzyme [23]. This fusion protein approach was further refined by introduction of a site-specific protease cleavage sequence between the p85a nSH2-iSH2 domains and the p110a protein. This fusion protein can be expressed in a soluble form and with high yields in insect cells using a recombinant baculovirus. After site specific cleavage during protein purification to release the p110a protein, a preparation is obtained that robustly yields high quality protein for co-crystallization with small molecule inhibitors [24]. Similarly, expression of soluble and active p110a protein requires co-expression of a p85a iSH2 construct in a baculovirus system. That system was used to prepare crystal grade protein to generate a series of high resolution complex structures with small molecule inhibitors in a program dedicated to the structure based discovery of isoform specific inhibitors [25].

1.6 Epigenetics Targets

Epigenetic processes are heritable states of gene expression that are not caused by changes in DNA sequences. Such mitotically and meiotically inheritable events have recently been demonstrated to be playing key roles in cancer genesis and tumor progression [26]. Epigenetic alterations, in contrast to genetic alterations, are of a reversible nature, at least in principle. This makes targeting of epigenetic processes attractive opportunities in small-molecule drug discovery. Examples for epigenetic processes are histone methylation and demethylation, histone acetylation and deacetylation among many others. Given the critical role of such post-translational modifications in the control of transcription, tight and reliable control of the respective enzyme activities are extremely important. It is therefore not surprising that very large protein machines execute epigenetic processes. These multi-protein complexes are inherently dynamic and individual components are able to assemble into very different complexes each having characteristic and distinct substrate specificities [26].

1.6.1 Histone Deacetylases

Histone deacetylases (HDACs) are a large family of enzymes that are responsible for the controlled and sequence-specific de-acetylation of nucleosomes. Lysine acetylation of histones H3 and H4 in nucleosomes is correlated with active (open) chromatin whereas deacetylation of these histones leads to compacted, and hence transcriptionally inactive, chromatin regions. Based on sequence analyses, eighteen different HDAC isoforms have been described in humans [27]. The core catalytic domain is generally contained in a region of about 300 amino acids in length. However, in the body, HDACs are never found in isolation. They rather are components of large multi-protein complexes that are formed by the temporal as well as the spatial recruitment of a large range of polypeptide cofactors as identified by studies aimed at elucidating the HDAC interactome [28, 29].

The catalytic domains of HDAC1 and HDAC2 for example are known to form heterodimers in the cell. These HDAC1/2 heterodimers are building the core of a number of different and defined multi-protein complexes with important and distinct cellular functions. Depending on cofactor composition the HDAC1/2 catalytic core can assemble into the NuRD complex, the Sin3 complex and into the CoREST complex, among others [29, 30].

The increasing insight into the importance of subunit composition for the function of epigenetic protein machines that are responsible for writing and erasing epigenetic marks, underscores the need to investigate such enzymes in their appropriate complexes in the context of well-defined subunit composition and stoichiometry. Small-molecule tool compounds can be immobilized to generate affinity matrices that can be used for pull-down experiments to identify interaction partners in cellular or tissue lysates [29, 30]. Alternative approaches to identify the exact subunit composition by proteomic techniques such as mass spectrometry open the way to rational recombinant expression of these complexes and to generate protein for *in vitro* assays and structure-based drug discovery. Recent data indicates for example that HDAC1 is tightly regulated within the NuRD complex. This regulation is in part mediated by the scaffolding component MTA1 that intimately interacts with the HDAC1 core by wrapping completely around the HDAC1 catalytic domain as visualized by the crystal structure of the HDAC1/MTA1 complex [31]. In the crystal structure, a binding site for the small molecule inositol-tetraphosphate [Ins(1,4,5,6)P_4] could be identified. This IP4 binding site has been originally discovered in the crystal structure of the related HDAC3/SMRT complex [32]. Subsequent *in vitro* activity assays clearly demonstrated that addition of IP4 to purified HDAC1/MTA1 and HDAC3/SMRT complexes strongly stimulates the HDAC enzyme activity. In accordance with the observation that the binding site for IP4 is formed by the interface that is created by protein/protein interactions between the two partners in the complex, the IP4 stimulation is only observed when the complexes are assayed *in vitro*. When tested in isolation, the HDAC subunits cannot be stimulated by IP4 addition and have been shown to have only very basal catalytic activity [31]. Both of these HDAC complexes were generated by transient expression in 293HEK cells. Since prokaryotes generally lack IP4, the important contribution of the small-molecule regulator would have probably missed out if a bacterial expression system had been employed for complex generation and structure solution. This dependence on availability of the appropriate small-molecule ligand in an expression host for the formation of a correctly folded and regulated protein complex will certainly be more important in the future as the discovery industry increasingly moves towards screening more “native-like” targets *in vitro*.

1.6.2 Protein Arginine Methyltransferases (PRMTs)

PRMTs are critical regulators of a number of vital cellular processes such as protein transport, regulation of gene expression and cellular signal transduction [26]. PRMT5 is the predominant enzyme responsible for arginine mono- and dimethylation [33]. *In vivo* PRMT5 is in a complex with the WD-repeat protein MEP50 that in turn is interacting with a range of cellular factors that confer substrate specificity and subcellular localization. Direct binding of MEP50 to PRMT5 greatly enhances the histone arginine methylation activity of PRMT5 predominantly by increasing the affinity of the enzyme complex towards its substrates. MEP50 essentially serves as a substrate recognition module for the catalytically active PRMT5 [33]. Active 1:1 protein complexes of full length PRMT5 and MEP50 can be readily produced by co-expression of the two genes in insect cells using recombinant baculoviruses. After purification the complex has marked histone arginine methyl-transferase activity *in vitro* and can be readily co-crystallized with small molecule inhibitors for structure-based drug design applications [34, 35]. The structures of PRMT5/MEP50 complexes reveal intricate interactions between the MEP β-propeller and the N-terminal domain of PRMT5 and support the functional analyses that indicated MEP50 to be an essential part of PRMT5 function.

1.7 The Ubiquitin-Proteasome System (UPS)

At a cellular level, one way of responding quickly to a change in environment is by rapid post-translational protein modification. Like protein phosphorylation, methylation, acetylation and citrullination, controlled protein degradation can be initiated much faster than it takes to manifest changes in the transcriptome [36]. In contrast to the reversible nature of protein phosphorylation, targeted degradation by the UPS pathway results in the irreversible removal of target proteins from the cellular context. The recent progress that has been made in our understanding of the role of the UPS in disease has triggered drug discovery programs aimed at identification and optimization of small molecule modulators of protein degradation processes. Tagging proteins with ubiquitin can reduce the half-life of proteins in extreme cases from several months to just a few minutes. Over 1000 different proteins are involved in the tagging process, many of which are attractive drug targets. The tagging proteins are grouped into three classes, called E1, E2 and E3. After E1 enzymes activate ubiquitin, E2 and E3 proteins attach it to their substrates. Once multiple ubiquitin moieties have been covalently attached to it, the tagged protein is then degraded by the proteasome. Deubiquitinating enzymes (DUBs) are proteases that can remove ubiquitin from proteins, adding another layer of regulation and complexity to the system [37]. Targeting the UPS with small molecules poses a number of challenges to the development of potent and selective small-molecule modulators. The chemistry of the enzymes involved is basically dependent on the presence of an active cysteine residue required for isopeptide bond formation. Active cysteine residues are therefore present in all E1, E2 and DUB proteins. Most small-molecule inhibitors that target these enzymes are therefore electrophiles and show only very limited selectivity. The currently most promising approach to selectively inhibiting the UPS appears to target the E3 subunits. E3 proteins mediate the interaction between the ubiquitin charged E2 and the substrate protein by directly interacting with both units and bringing them in close proximity. E3 proteins represent the largest family of subunits within the UPS pathway since they are directly involved in target recognition and therefore offer the potential to be amenable to the development of potent and selective inhibitors. However, E3 proteins are basically adaptor molecules that mediate protein-protein interactions (PPIs) and do not have an enzyme activity themselves that could be inhibited. Small molecule inhibitors targeting E3 proteins therefore are inhibitors of protein-protein interactions. PPI inhibitors, however, are generally more difficult to find and optimize. The development of PPI inhibitors benefits massively from structural information and therefore the UPS pathway has been the focus of intense efforts to elucidate the structural details of interactions of ubiquitin ligases in complex with their substrates [38]. Generally these complexes are assembled *in vitro* after expression and purification of the individual chains in *E. coli*. Increasingly, such protein complexes are used as the starting point for fragment-based drug discovery programs, with the aim of finding new small-molecule binding sites that can be used for selective inhibitor design. Finding fragments and determination of their binding site and modes enables the medicinal chemists to start from novel chemistry. That method is better suited to target PPIs, also because the enzyme inhibitor compounds that are present in large historical compound libraries have been designed and optimized for very different target classes [37].

While the majority of protein complexes are being generated using engineered expression systems, there is still a need for purified proteins from native sources to support small-molecule drug discovery. The proteasome provides an interesting example of a challenging multi-protein complex that so far cannot be generated by recombinant expression systems. The proteasome is a multi-subunit protein complex that contains a large number of different subunits. For correct and tightly controlled assembly of this large complex, a number of assembly factors are transiently needed in the cell. They have to act in a concerted and exactly timed fashion in order to assemble functional proteasomes with the correct subunit composition [39, 40]. The complexities faced by the number of different protein cofactors, which are required to form and assemble proteasomes, has so far prevented their successful recombinant production. Inhibitor studies and structural biology therefore relies on proteasomes purified from their native sources. Most of the work done in the past relied on the yeast 20S proteasome that can readily be extracted from commercially available baker's yeast. The purified protein is active *in vitro* and can be readily crystallized with a number of small-molecule inhibitors that were designed to inhibit the human proteasome. A large number of complex structures have been solved this way, supporting the claim that the yeast 20S proteasome can be in many cases a suitable surrogate for the more complex human version [41]. Recently, the first human 20S proteasome structures have been reported [42]. The structures were obtained after the human 20S proteasome complex was purified from human erythrocytes, a readily available starting material that could be obtained from a blood center.

1.8 Conclusion

The examples listed above can only offer a glimpse of the many and multi-faceted ways that protein complexes are increasingly being used by the drug discovery industry for the development of novel therapeutics. Too many costly failures of drug candidates in the clinic, either due to lack of efficacy or adverse side effects, have prompted a paradigm shift in the industry. Increasingly target-based drug discovery campaigns are focused on protein starting material that as closely as possible resembles the target situation *in vivo*. The need to boost drug discovery efficiency and cost structure is being fueled by our ever-increasing understanding of the structure and function of the many multi-protein machines in our cells. As our understanding of the importance to study protein complexes in their correctly regulated states becomes more complete, the more pressing the need for efficient and economic expression systems gets. It is now generally acknowledged that for drug discovery protein complexes are the more relevant reagents. However, the move away from the study of catalytic domains in isolation to more relevant protein complexes *in vitro* not only requires continuous refinement of expression systems. Complexes observed *in vivo* need to be thoroughly characterized by state of the art techniques with respect to minimal subunit composition, exact stoichiometry and post-translational modification patterns. The best protein complex characterization achievable does have a dramatic impact on the chances of success in target based drug discovery.

References

1. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92:291–294
2. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlasblom J, Dar VU, Bezginov A, Clark GW, Wu GC, Wodak SJ, Tillier ER, Paccanaro A, Marcotte EM, Emili A (2012) A census of human soluble protein complexes. Cell 150:1068–1081
3. Larance M, Lamond AI (2015) Multidimensional proteomics for cell biology. Nat Rev Mol Cell Biol 16:269–280
4. Assenberg R, Wan PT, Geisse S, Mayr LM (2013) Advances in recombinant protein expression for use in pharmaceutical research. Curr Opin Struct Biol 23:393–402

5. Manning G, Whyte DB, Martinez R et al (2002) The protein kinase complement of the human genome. Science 298:1912–1934
6. Malumbres M (2014) Cyclin-dependent kinases. Genome Biol 15:1–10. doi: 10.1186/gb4184
7. Echalier A, Endicott JA, Noble MEM (2010) Recent developments in cyclin-dependent kinase biochemical and structural studies. Biochim Biophys Acta 1804:511–519
8. Jeffrey PD, Russo AA, Polyak K et al (1995) Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. Nature 376:313–320
9. Russo AA, Jeffrey PD, Pavletich NP (1996) Structural basis of cyclin-dependent kinase activation by phosphorylation. Nat Struct Biol 3:696–700
10. Day PJ, Cleasby A, Tickle IJ et al (2009) Crystal structure of human CDK4 in complex with a D-type cyclin. Proc Natl Acad Sci U S A 106:4166–4170
11. Schulze-Gahmen U, Kim S-H (2002) Structural basis for CDK6 activation by a virus-encoded cyclin. Nat Struct Biol 9:177–181
12. Baumli S, Lolli G, Lowe ED et al (2008) The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation. EMBO J 27:1907–1918
13. Sken CABO, Farnung L, Hintermair C et al (1AD) The structure and substrate specificity of human Cdk12/Cyclin K. Nat Commun 5:1–14
14. Grahame Hardie D (2014) AMP-activated protein kinase: a key regulator of energy balance with many roles in human disease. J Intern Med 276:543–559
15. Rana S, Blowers EC, Natarajan A (2015) Small molecule adenosine 5′-monophosphate activated protein kinase (AMPK) modulators and human diseases. J Med Chem 58:2–29
16. Neumann D, Woods A, Carling D et al (2003) Mammalian AMP-activated protein kinase: functional, heterotrimeric complexes by co-expression of subunits in Escherichia coli. Protein Expr Purif 30:230–237
17. Rajamohan F, Harris MS, Frisbie RK et al (2010) Escherichia coli expression, purification and characterization of functional full-length recombinant alpha-2beta2gamma3 heterotrimeric complex of human AMP-activated protein kinase. Protein Expr Purif 73:189–197
18. Xiao B, Sanders MJ, Underwood E et al (2011) Structure of mammalian AMPK and its regulation by ADP. Nature 472:230–233
19. Xiao B, Sanders MJ, Carmena D, et al (1AD) Structural basis of AMPK regulation by small molecule activators. Nat Commun 4:1–10
20. Thorpe LM, Yuzugullu H, Zhao JJ (2015) PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. Nat Rev Cancer 15:7–24
21. Dobbelstein M, Moll U (2014) Targeting tumour-supportive cellular machineries in anticancer drug development. Nat Rev Drug Discov 13:179–196
22. Huang C-H, Mandelker D, Schmidt-Kittler O et al (2007) The structure of a human p110alpha/p85alpha complex elucidates the effects of oncogenic PI3Kalpha mutations. Science 318:1744–1748. doi:10.1126/science.1150799
23. Hu Q, Klippel A, Muslin AJ et al (1995) Ras-dependent induction of cellular responses by constitutively active phosphatidylinositol-3 kinase. Science 268:100–102
24. Sinnamon RH, McDevitt P, Pietrak BL et al (2010) Baculovirus production of fully-active phosphoinositide 3-kinase alpha as a p85alpha-p110alpha fusion for X-ray crystallographic analysis with ATP competitive enzyme inhibitors. Protein Expr Purif 73:167–176
25. Berndt A, Miller S, Williams O et al (2010) The p110δ structure: mechanisms for selectivity and potency of new PI(3)K inhibitors. Nat Chem Biol 6:117–124
26. Yoo CB, Jones PA (2006) Epigenetic therapy of cancer: past, present and future. Nat Rev Drug Discov 5:37–50
27. Micelli C, Rastelli G (2015) Histone deacetylases: structural determinants of inhibitor selectivity. Drug Discov Today 20:718–735
28. Joshi P, Greco TM, Guise AJ et al (2013) The functional interactome landscape of the human histone deacetylase family. Mol Syst Biol 9:1–21
29. Bantscheff M, Hopf C, Savitski MM et al (2011) Chemoproteomics profiling of HDAC. Nat Biotechnol 29:255–265
30. Drewes G (2012) Future strategies in epigenetic drug discovery. Drug Discov Today Ther Strateg 9:e121–e127
31. Millard CJ, Watson PJ, Celardo I et al (2013) Class I HDACs share a common mechanism of regulation by inositol phosphates. Mol Cell 51:57–67
32. Watson PJ, Fairall L, Santos GM, Schwabe JWR (2012) Structure of HDAC3 bound to co-repressor and inositol tetraphosphate. Nature 481:1–7
33. Stopa N, Krebs JE, Shechter D (2015) The PRMT5 arginine methyltransferase: many roles in development, cancer and beyond. Cell Mol Life Sci 72:2041–2059
34. Antonysamy S, Bonday Z, Campbell RM et al (2012) Crystal structure of the human PRMT5:MEP50 complex. Proc Natl Acad Sci U S A 109:17960–17965
35. Chan-Penebre E, Kuplast KG, Majer CR et al (2015) A selective inhibitor of PRMT5 with in vivo and in vitro potency in MCL models. Nat Chem Biol 11:432–437
36. Nalepa G, Rolfe M, Harper JW (2006) Drug discovery in the ubiquitin–proteasome system. Nat Publ Group 5:596–613
37. King RW, Finley D (2014) Sculpting the proteome with small molecules. Nat Chem Biol 10:870–874
38. Gaczynska M, Osmulski PA (2015) Targeting protein-protein interactions in the proteasome super-assemblies. Curr Top Med Chem 15(20):2056–2067
39. Gallastegui N, Groll M (2010) The 26S proteasome: assembly and function of a destructive machine. Trends Biochem Sci 35:634–642. doi:10.1016/j.tibs.2010.05.005
40. Gu ZC, Enenkel C (2014) Proteasome assembly. Cell Mol Life Sci 71:4729–4745

41. Huber EM, Heinemeyer W, Groll M (2015) Bortezomib-resistant mutant proteasomes: structural and biochemical evaluation with Carfilzomib and ONX 0914. Struct Fold Des 23:407–417

42. Harshbarger W, Miller C, Diedrich C, Sacchettini J (2015) Crystal structure of the human 20S proteasome in complex with Carfilzomib. Struct Fold Des 23:418–424

Choose a Suitable Expression Host: A Survey of Available Protein Production Platforms

2

Francisco J. Fernández and M. Cristina Vega

Abstract

Recombinant overexpression of a protein or a protein complex using any specific heterologous host can be an overwhelming challenge. The reasons may range from low yield and poor solubility of a single-subunit enzyme to the wrong stoichiometry or the incomplete assembly of a multiprotein complex. Whatever the reason, overcoming the difficulties will take the researcher into a journey through the seemingly countless options that exist for protein expression. While some choices stand to reason fairly straightforwardly (*e.g.*, using *Escherichia coli* for the production of bacterial enzymes), most other choices do not need to be so self-revealing. Here, we attempt to portrait the canvas of available hosts for heterologous expression of many different protein classes and complexes and offer guidance as to which expression host may be more suitable to the problem at hand. The guidance in this chapter must be taken only as a rough indication which will have to be checked against the available literature and corroborated by experiment. It is not only expected but also welcome that, as more knowledge is gathered about the performance of hosts and protein types and new expression systems develop, the information in this chapter will have to be updated and refined.

Keywords

Heterologous expression • Protein complex • Expression host • *E. coli* • Yeast • Baculovirus • Mammalian culture

2.1 *E. coli* as the Workhorse for Recombinant Expression

Since the advent of molecular biology that made possible the use of heterologous hosts for the overexpression of proteins and protein complexes,

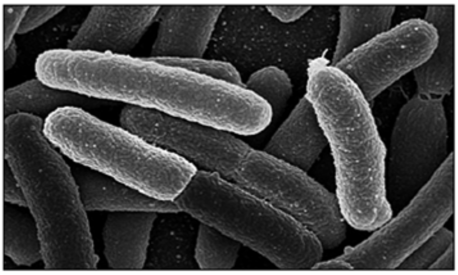
F.J. Fernández (✉) • M.C. Vega (✉)
Center for Biological Research, Spanish National Research Council (CIB-CSIC),
Ramiro de Maeztu 9, 28040 Madrid, Spain
e-mail: cvega@cib.csic.es

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_2

a tremendous wealth of knowledge has accumulated on the subject and, in particular, in using the enterobacterium *Escherichia coli* for the production of recombinant proteins (Fig. 2.1). Several excellent reviews have been recently published highlighting the usefulness, breadth of applicability, and straightforward use of *E. coli*, *e.g.*, see Rosano and Ceccarelli [1]. Most of the advantages of *E. coli* for protein production are well established: it has extremely fast growth kinetics, high cell densities can be achieved, culture media and reagents are inexpensive, and transformation with expression constructs is straightforward. For the structural biology and chemical biology fields, the ease with which labeled or non-natural amino acids can be incorporated in *E. coli* expression cultures is a definitive advantage; important examples include the incorporation of selenomethionine and selenocysteine for X-ray crystallography [2] or isotopic labeling for nuclear magnetic resonance applications [3]. An additional advantage has already been mentioned that cannot be emphasized enough: a wealth of experimental knowledge and successful test cases spanning over many different proteins and complexes.

Useful generalizations have been drawn from the accumulated knowledge. For example, we now know that single-subunit metabolic enzymes of moderate size and typical isoelectric point (pI) bearing hydrolytic activities might be overexpressed conveniently in *E. coli*. A vast literature argues that a bewildering amount of protein classes and sizes can be obtained from this amazing bacterium. Conversely, some protein types are known to pose a considerable challenge to *E. coli*: from protein kinases (which tend to kill growing *E. coli* cells) to large eukaryotic multisubunit complexes (which typically lead to no expression or very poor yields) and membrane proteins, many of which require specific post-translational modifications (PTMs) and lipid compositions. Some structural genomics consortia have used *E. coli* as their expression workhorse with excellent results, demonstrating that very broad protein families, prokaryotic and eukaryotic, can be expressed in large amounts, in soluble and functional form, in *E. coli*. Connected with this predominant use of *E. coli* for recombinant protein expression and the large-scale projects that have used it systematically, there is a wealth of tools and of ongoing development efforts to create new tools that push *E. coli* expression systems to perform better in those areas where it traditionally does badly. It is no surprise that the majority of the crystal structures deposited at the Protein Data Bank (PDB) come from pure and homogeneous material purified from overexpressing *E. coli* cells.

In light of the above, the first piece of advice is easy to understand: unless the protein or complex (or protein class or complex class) to be overproduced is known to behave better in another system, always use *E. coli* first. Optimization of the expression construct is nearly always necessary in *E. coli*, including, but not limited to, choice of the nature and placement of fusion partners (N or C-terminal, small peptide



Escherichia coli, the workhorse of protein expression

>70% of all recombinant proteins (PubMed) (2013)
>88% of all proteins in the Protein Data Bank (2015)

Fig. 2.1 ***E. coli as the workhorse of protein expression***. The impressive record of *E. coli* in the protein expression arena can be quantitated by the percentage of all recombinantly expressed proteins that have been produced in *E. coli* as judged by a standard literature search (PubMed) and from the protein structure databases (in particular, the PDB). The latter is of prime importance, since most structural characterization requires samples of the highest possible quality

tags or larger fusion proteins), truncation (from either end or both ends), truncation of loops, and mutagenesis (*e.g.*, to increase thermostability, to reduce surface entropy).

E. coli also offers a large repertoire of possibilities as a host for protein coexpression, which significantly widens the range of proteins and complexes that can be tackled [4]. The ACEMBL system, for example, exploits the concept of tandem recombineering and Cre-loxP recombination to enable researchers to quickly build many multigene constructs for their expression [5]. An up-to-date review of ACEMBL is presented in Chap. 3. Other approaches for coexpression in *E. coli* have been put forward, especially in the context of high-throughput approaches where quickly reviewing many coexpression experiments to pinpoint the variables that are most crucial for success, such as genes, truncations, and tag identity and position, becomes the focus of research. Some of those approaches are discussed in Chap. 4.

A specialized field where *E. coli* has been very useful is in the expression of membrane proteins (reviewed in Chap. 5). There are numerous examples of successfully expressed membrane proteins from prokaryotic and eukaryotic origin alike. In part at least, this remarkable success is due to the ease with which the *E. coli* toolbox can be optimized to tackle very challenging tasks. Production of functional membrane proteins in *E. coli* required the isolation of cell strains with more extensive membrane systems and highly tolerant to the accumulation of toxic proteins, devising innovative optimization and stabilization strategies, creating downstream analytic techniques to probe into the fold state of membrane proteins via fluorescent tags, and experimenting with various fusion partners.

Despite its usefulness, *E. coli* has limitations as an expression host and it cannot be assumed that it will yield useful amounts of recombinant protein in general. This is especially so when it comes to the production of proteins for which little information is available (*e.g.*, functional screening from metagenome libraries [6]) or when the target consists of large eukaryotic proteins and complexes [2]. There are other inherent drawbacks of using *E. coli* for all expression experiments including inclusion body formation (although sometimes inclusion bodies can be advantageous [7]), lack of stability of large multigene constructs over many generations, differences in codon usage to the organism where the target gene originates, toxicity, membrane structure and composition, failure to properly fold and/or assemble proteins and complexes, lack of organelles, lack or an insufficient supply of cofactors, etc.

When a reasonably large set of expression experiments has been conducted in *E. coli* (including one with full-length versions of the genes of interest) without promising results, the next best strategy is to move on to one of the possible alternative expression hosts.

2.2 Beyond *E. coli*

As it was pointed out in the previous section, as the difficulty to express the gene of interest (GOI) increases—measured operatively by a general failure to obtain soluble, active protein (complex) after performing a reasonably large number of more conventional expression tests in *E. coli*—, other expression hosts must be considered. Searching for the next expression host beyond *E. coli* is not an easy task in general since there is no expression system that can guarantee success for any arbitrary protein or protein complex, and because the decision has to include aspects of the biology of the system under study and more pragmatic considerations such as available infrastructure and previous experiences.

There are many prokaryotic hosts that offer similar advantages to *E. coli* in terms of speed, convenience, and cost-efficiency, while adding extra versatility not found in *E. coli*. For example, *Bacillus megaterium* can secrete proteins that would be hard for *E. coli* to secrete—including many enzymes of moderate size. In particular, if the end product is for therapeutic use or for human consumption, *B. megaterium* becomes an attractive alternative to *E. coli* for its production even as intracellular protein because of its GRAS (generally regarded as safe) status. The use of *B.*

megaterium as a host for heterologous expression is discussed in Chap. 7.

Other bacteria have been researched as heterologous hosts, most of the time as tailor-made solutions for specific proteins that were deemed to possess very specific features that made them recalcitrant for *E. coli* expression. Chapter 8 deals with some of the most representative groups that have been used as alternative bacterial expression hosts: *Pseudomonas*, *Streptomyces*, and, to a lesser extent, *Mycobacteria*. In addition to the latter, Chap. 8 also summarizes the use of other prokaryotic hosts, this time from the Archaea domain, for heterologous expression. *Pseudomonas* has been used for the production of certain oxidoreductases, and the success of this expression system has been attributed to their strictly aerobic metabolism—which would make them ideal for the proper folding and stabilization of the recombinant oxidoreductases. The mycelial soil bacterium *Streptomyces lividans*, for example, has attracted attention as a producer of secondary metabolites and for the production of drug-modifying enzymes. In fact, this Actinobacterium accounts for more than half of antibiotics production in the market. Mycobacteria are also interesting hosts for overproduction of mycobacterial proteins, some of which are very specific to the mycobacterial hosts, *e.g.*, iron-superoxide dismutase and cell-wall and specific lipid biosynthetic enzymes. The use of very specific hosts for the production of very specific protein targets, which might be very hard to produce in generalist expression systems such as *E. coli*, represents indeed a leit motif in the field: use generalist systems whenever possible but switch to a specialty system when required.

Archaeal systems have been tested in a few specialized cases, which are discussed in Chap. 8. The archetypical example is *Halobacterium salinarum* for the production of the light-harvesting bacteriorhodopsin, a membrane protein that clutters the wild-type archaeon's membrane endowing it with its typical red color. This observation has motivated authors to suggest that this archaeal system could be able to express mammalian GPCRs, a suggestion that still needs to be fully checked after a first success in the overexpression of the human G-protein coupled β2-adrenergic receptor.

Chapter 9 presents an overview of the current understanding on yeasts as expression hosts. Yeasts share many useful features with *E. coli* and other bacterial hosts as expression hosts, including fast growth rates, inexpensive culture media, many useful molecular biology tools (including, *e.g.*, promoters, selection markers) and very accessible genomes for targeted genetic manipulation. These properties, combined with the eukaryotic nature of yeasts, which allow them to perform many PTMs and to provide sophisticated folding machinery for the efficient production of eukaryotic protein machines, have spurred the idea that yeasts should be used more frequently as the first alternative microorganism to *E. coli* [8]. For eukaryotic proteins of fungal origin, some naturally located in the cytoplasm (lactase, lipases) or secreted to the extracellular medium (glycosidases, peptidases), unicellular fungi (yeasts) or mycelial fungi (filamentous fungi) might be more appropriate hosts for expression, especially given the more efficient secretory pathways of yeasts and fungi when compared with prokaryotic hosts. Given its secondary role in protein expression, filamentous fungi are dealt with in Chap. 11 together with *Dictyostelium discoideum*. Many eukaryotic (including human) proteins and peptide hormones that are typically secreted can usually be made using yeasts and fungi, including many protein factors from the immune complement system. When it comes to yeasts, there is one important choice to make between methylotrophic and non-methylotrophic yeasts. The latter, including *S. cerevisiae*, *K. lactis* and *Y. lipolytica*, are very attractive because of their better-known genetics and metabolism and because they can be engineered for rapid protein production in screening and high-throughput settings. In contrast, methylotrophic yeasts (prominently, *K. pastoris* and *O. polymorpha*) grow to higher densities and typically produce greater yields for most proteins at the expense of a greater investment in inserting the expression cassette into the genome and screen for correct, expressing transformants. Together, the two types of yeasts can be combined in a powerful combo, with non-methylotrophs used for fast screening and optimization purposes and methylotrophs employed mostly for the generation of industrial-scale overproducing strains and to boost protein product yields.

Other lower eukaryotic hosts to be considered are protists. *Leishmania tarentolae* (described in Chap. 10), for example, has been shown to be particularly effective for the production of kinases, membrane proteins and Cu/Zn superoxide dismutase, as well as several membrane proteins. An attractive feature of *L. tarentolae* is that the expression cassettes can be inserted into its genome with ease, yielding cell lines that can overexpress several protein chains simultaneously.

Dictyostelium discoideum (discussed in Chap. 11 with filamentous fungi) has excelled for the production of cytoskeletal proteins, perhaps owing to the specializations of *D. discoideum* for a highly active, motile lifestyle. The social amoeba has the enzymatic machinery to decorate glycoproteins with nearly mammalian glycosylation patterns, a property that becomes interesting for pharmacological target proteins as an alternative to highly engineered mammalian glycosylation mimicking systems.

Many eukaryotic proteins and protein complexes, however, have special requirements in terms of PTMs, chaperone assistance, and folding properties, that require insect or mammalian cells for their correct folding, processing and assembly. These systems are more complex to handle, require more training and are relatively more expensive, although advances in the fields are making them ever more accessible and cost-effective. Chapters 12 and 13 discuss recent advances in insect cell expression, while Chaps. 14 and 15 cover state-of-the-art in mammalian expression systems. Particularly noteworthy is the availability of automated systems for protein coexpression for insect cells and mammalian cells, *e.g.*, MultiBac [9] and MultiLabel [10], which extend their usefulness precisely to the targets most difficult to express in *E. coli* and other systems.

Other systems do exist that harbor distinct advantages. Among them, cell-free systems are attractive because, once available and set up, they do not require growth of cells nor lysis and extraction, and they can be optimized to express large amounts of proteins with minimal impurities. Chapter 6 deals with cell-free systems, and Chap. 10 comments on the use of the *L. tarentolae* expression system as a source for cell-free extracts competent for transcription-translation experiments.

Plants and algae have shown promise for several types of proteins, including plant enzymes, proteins, and sophisticated light-harvesting protein complexes. New and exciting techniques are emerging that allow the introduction of very complex expression cassettes, potentially harboring tens of different genes, into plant cells, and the maturation of those techniques holds the promise of making plant hosts more attractive for the recombinant production of protein complexes and the assembly of very complex enzymatic pathways. These should be further refined to reach a higher degree of performance and to become more widespread in use. They are treated in Chaps. 16, 17, and 18.

Multiprotein complexes can be obtained also from its constituent protein chains if they can be expressed independently. For those favorable cases where this is possible, complex reconstitution offers a straightforward route to test many variables, assemble the complex in several stoichiometries, pre-add drugs or small-molecule inhibitors to some of the polypeptide chains before complexation, and other approaches. Chapter 19 gives a glimpse into this vast field, and Chap. 20 explains advances made in the understanding of intracellular signaling cascades that have been partly facilitated by complex reconstitution strategies.

Today one of the frontiers for protein expression is the production of membrane proteins. The need to recreate the delicate and specific hydrophobic environment where membrane proteins function pose many experimental challenges, including problems of membrane protein localization, membrane insertion, folding, PTMs, and more. Tackling these problems requires access to appropriate protocols and reagents. In Chap. 5, an overview is given (centered in, but not necessarily limited to, *E. coli*), the general requirements of membrane proteins are described and some of the possibilities discussed, in particular for expressing multisubunit membrane proteins, a feat that is actively researched today and is bound to grow in relevance, as more membrane proteins become the target of future investigations.

Properly assessing whether (and to what an extent) a multisubunit complex has been assembled following a coexpression or a reconstitution experiment involves a variety of techniques and approaches. In Chaps. 21 and 22 several powerful methodologies are presented that enable the acquisition of valuable information on the interaction, stability or transient nature of a complex, its shape and size, and the stoichiometry. Since some of these methodologies are quite advanced, those chapters list additional literature that should provide further background material.

2.3 Features of the Target Gene/s

Before attempting to select an expression host, the features of the gene or genes to express should be considered. Both bioinformatics tools and available experimental knowledge, either from the literature or produced in the laboratory, should be thoroughly examined. Among those features one should consider the use of full-length genes or truncated constructs encoding one or several domains only, the overall size of the construct, and the %G+C.

The temptation to generate many truncated constructs should be contained at the onset of a protein expression experiment, where it is more important to gain information as to the solubility and correct folding of the target protein. At a later stage, the exact point of truncation at the N- and C-terminal ends can be screened to either improve upon a minimum level of soluble expressed protein or as a procedure to generate soluble protein when none was obtained from previous constructs—this approach has been exploited for high-throughput large-scale screening protocols [11, 12].

Construct size is an important consideration, because *E. coli* cannot really express proteins larger than 120 kDa very efficiently, and those are typically expressed in very low yields, targeted to inclusion bodies or extensively proteolytically degraded (although there is anecdotal evidence than proteins about 300 kDa in size have been overexpressed in *E. coli*).

The identity of the first nucleotides of the coding sequence and the secondary and tertiary structure of the mRNA should also be carefully reviewed, since they can have a great impact on the transcription rates and hence in the overall protein expression yield.

Extremely high or low %G+C genes should be either expressed in the native source (if the target gene is known to express to high levels naturally) or modified, typically by synthesizing a custom gene encoding the same protein, so that it contains a more balanced %G+C. In A+T rich genes it is not uncommon to find cryptic stalling sites for the ribosome, which ultimately lead to mRNA truncation and apparently "proteolyzed" proteins.

Protein features need to be considered as well. For example, highly active oxidoreductases or kinases may require special properties from the host to avoid toxicity problems. Should PTMs be absolutely required for proper folding, these requirements have to be procured for a successful production, either in the selection of an appropriate host or by coexpression with the modifying enzymes. Proteins that are targeted for secretion will in general require hosts with an efficient secretory pathway, such as yeasts or filamentous fungi.

Finally, protein intracellular localization is a strong constraint for many proteins, including membrane proteins. Proteins that function in specific eukaryotic organelles (*e.g.*, mitochondria, chloroplasts, peroxisomes) will generally benefit from expression hosts that possess such organelles.

When a protein complex is the target of an expression experiment, other important features to pay attention to are: which proteins form the stable core of the complex (if there is any), which proteins attach to this core transiently, which are the factors or modifications that trigger complex assembly when the latter is a regulated process, whether complexation leads to stabilization of certain protein sequences that were unstructured before, etc. All these factors are important and exclusively considering the individual proteins may not be informative, since the simultaneous translation of several "unstable" proteins might in fact yield a "stable" complex when and/or if the floppy parts of those proteins are engaged in a stable interaction within the complex. A beautiful example of this is the structure of the spliceosome [13]. The bottom line is that a host should be chosen that provides all the essential requirements for the protein or protein complex to be produced, or a host must be supplemented with the heterologous genes that provide such requirements.

2.4 Generalist Versus Specialists

A useful difference is made between those expression hosts that have been proven successful to overexpress broad classes of target proteins, like *E. coli*, *K. pastoris* and baculovirus-infected insect cells, which we call "generalists", and those expression hosts that have been used in specific cases only, which we call "specialists" [14]. Generalist hosts are especially recommended as the first expression systems to try on proteins and protein complexes from any source, for uncomplicated proteins and protein complexes, or when little prior knowledge is available about the target proteins. In contrast, a specialist host may be better suited for the task when the target protein is known to have been successfully overexpressed only in a particular host, or when the specific requirements for the target protein are known (or suspected) to be matched only by one or a few expression hosts. An example of the latter case is the industrial production of antibodies in carefully optimized mammalian cell culture—in this system, all the specific requirements necessary for high-titer expression of antibodies are matched, and the transcription, translation and secretory machinery of the cells are specifically fine-tuned for antibodies. Attempts to express antibodies in many other expression systems have been met with varying degrees of success, but even in those cases where antibodies can be expressed and secreted in large amounts, rarely do they come even close to the productivities reached by highly optimized mammalian cell cultures.

The native host may be considered to be an extreme example of a "specialized" host—except when the native host is one of the generalist expression hosts, *e.g.*, *E. coli*. Sometimes the native host is the best system to express a given protein, even under the control of a heterologous promoter, but this cannot be taken for granted; sometimes, the native host may turn out to be the worst host when overexpression leads to severe toxicity. Success of the native host depends on the specific protein (complex) construct. For example, fungal extracellular enzymes might be best produced in a fungal expression system; but a yeast enzyme with a lethal mutation in the gene's coding sequence might lead to reduced growth and poor yields if expressed in its native or a related host. Expressing a lethal mutant protein in a heterologous host, where the chances that the defective protein may interfere with the normal growth and function of the heterologous host are reduced, might indeed be the most sensible action.

2.5 Recommendations

Although we admit that giving strong recommendations as to which expression host to use for a particular protein or protein complex is intrinsically problematic, we should not refrain from constructing some useful suggestions (summarized graphically in Fig. 2.2). This recommendation should be viewed as a rough guide in the absence of more specific information about the target protein or complex. In particular, we have constructed this set of recommendations based on the microbial and cellular expression hosts described in this book; other expression hosts do exist that may be better suited for the production of specific proteins. It should be kept in mind that the same results might also be obtained using a variety of host and plasmids combinations.

Obviously, when more specific information is available, these recommendations should be updated to reflect the new information.

E. coli (Chaps. 3 and 4) should nearly always be used in a first attempt, since when it works it provides a fast and cheap route to the expressed target. Many strains and plasmids are available to introduce several PTMs, to enable the secretion of expressed proteins (especially when they are small), to assist the folding of disulfide-bridge containing proteins, and to assemble coexpression constructs. Chapter 5 describes *E. coli* as a host for the overexpression of membrane proteins and Chap. 6 discusses the use of cell-free extracts for protein expression.

B. megaterium (Chap. 7) is a better secretor than *E. coli* and has been used for the production of virulence factors.

Pseudomonas (Chap. 8) has been used to produce fimbrial proteins, endogenous receptors, heme-containing proteins, and high % G+C endonucleases.

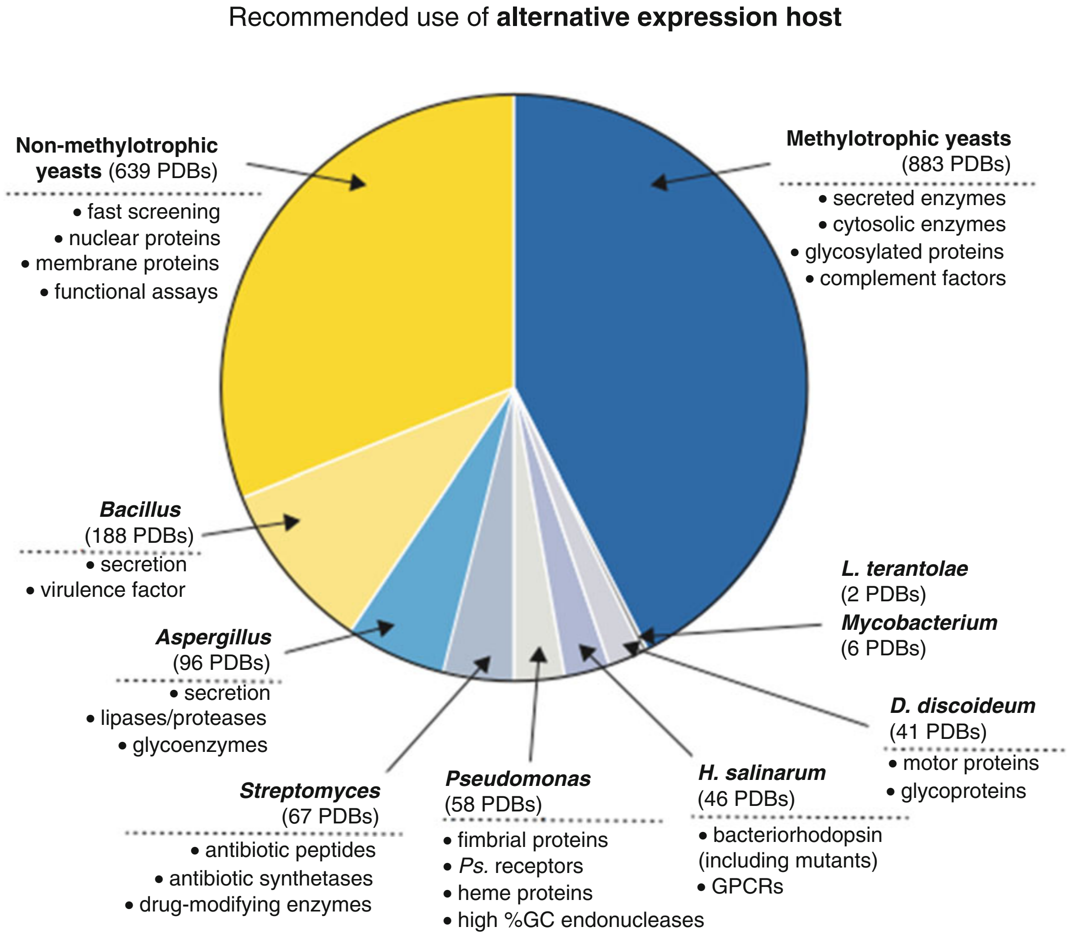


Fig. 2.2 Selection of a suitable alternative expression host for a protein expression project. Pie chart showing several alternative expression hosts, with section areas representing frequency of use in the PDB. To aid in choosing from them, guidelines are shown underneath each expression host regarding the recommended use for that host (with kind permission from Elsevier [14])

Streptomyces (Chap. 8) has outperformed other hosts in the production of antibiotic peptides, antibiotic synthetases and drug-modifying enzymes.

Mycobacterium (Chap. 8) is used for very specific mycobacterial proteins that might require excessive strain engineering in a heterologous host.

Halobacterium and other halophilic archaeon (Chap. 8) have been utilized for the production of the bacteriorhodopsin membrane protein.

Non-methylotrophic yeasts (Chap. 9) as *S. cerevisiae*, *K. lactis* and *Y. lipolytica* have been used quite successfully for fast screening of many coding sequences, nuclear proteins, membrane proteins, functional assays, and for the secretion of many useful enzymes.

Methylotrophic yeasts (Chap. 9) as *K. pastoris* and *O. polymorpha* have use for the production of secreted enzymes, cytosolic enzymes, glycosylated proteins and complement factors.

Leishmania tarentolae (Chap. 10) has its area of application in kinases and superoxide dismutases.

Dictyostelium discoideum (Chap. 11) is particularly well suited for the production of motor/cytoskeletal proteins and near-mammalian glycosylated proteins.

Filamentous fungi (Chap. 11) such as *Aspergillus* spp. and *Trichoderma reesei* have very efficient secretory pathways and have been used for the production of secreted enzymes, lipases, proteases and glycoenzymes.

2.6 Conclusions

The choice of host for the production of specific proteins and/or protein complexes influences the outcome of an expression experiment, therefore care should be paid when making this early and crucial choice. There are many expression hosts available, and it is likely that more hosts (especially, microbial hosts) will be characterized over the coming years. The ideal host must be determined for each expression target, although general recommendations can be drawn. A useful simplification divides all expression hosts in “generalists”, which can be applied successfully to a wide variety of problems, are well characterized and studied, and “specialists”, which might have specific properties useful for the case at hand. Among the generalists the most frequently used host is *E. coli*, but other microbial and cellular hosts are available: *S. cerevisiae*, *K. pastoris*, insect and mammalian cells. The specialists include numerous (micro-)organisms and their use reflects the idiosyncrasies of the proteins to be expressed.

Acknowledgments This work was funded by the Spanish Ministry of Science and Innovation grant CTQ2015-66206-C2-2-R and the European Commission (Framework Programme 7) (FP7) project ComplexINC No. 279039 to M.C.V.

References

1. Rosano GL, Ceccarelli EA (2014) Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol 5:172
2. Strub MP, Hoh F, Sanchez JF, Strub JM, Bock A, Aumelas A, Dumas C (2003) Selenomethionine and selenocysteine double labeling strategy for crystallographic phasing. Structure 11(11):1359–1367
3. Cai M, Huang Y, Sakaguchi K, Clore GM, Gronenborn AM, Craigie R (1998) An efficient and cost-effective isotope labeling protocol for proteins expressed in Escherichia coli. J Biomol NMR 11(1):97–102
4. Hochkoeppler A (2013) Expanding the landscape of recombinant protein production in Escherichia coli. Biotechnol Lett 35(12):1971–1981
5. Bieniossek C, Nie Y, Frey D, Olieric N, Schaffitzel C, Collinson I, Romier C, Berger P, Richmond TJ, Steinmetz MO, Berger I (2009) Automated unrestricted multigene recombineering for multiprotein complex production. Nat Methods 6(6):447–450
6. Liebl W, Angelov A, Juergensen J, Chow J, Loeschcke A, Drepper T, Classen T, Pietruszka J, Ehrenreich A, Streit WR, Jaeger KE (2014) Alternative hosts for functional (meta)genome analysis. Appl Microbiol Biotechnol 98(19):8099–8109
7. Hwang PM, Pan JS, Sykes BD (2014) Targeted expression, purification, and cleavage of fusion proteins from inclusion bodies in Escherichia coli. FEBS Lett 588(2):247–252
8. Bill RM (2014) Playing catch-up with Escherichia coli: using yeast to increase success rates in recombinant protein production experiments. Front Microbiol 5:85

9. Bieniossek C, Imasaki T, Takagi Y, Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 37(2):49–57
10. Kriz A, Schmid K, Baumgartner N, Ziegler U, Berger I, Ballmer-Hofer K, Berger P (2010) A plasmid-based multigene expression system for mammalian cells. Nat Commun 1:120
11. An Y, Yumerefendi H, Mas PJ, Chesneau A, Hart DJ (2011) ORF-selector ESPRIT: a second generation library screen for soluble protein expression employing precise open reading frame selection. J Struct Biol 175(2):189–197
12. Yumerefendi H, Tarendeau F, Mas PJ, Hart DJ (2010) ESPRIT: an automated, library-based method for mapping and soluble expression of protein domains from challenging targets. J Struct Biol 172(1):66–74
13. Nie Y, Viola C, Bieniossek C, Trowitzsch S, Vijay-Achandran LS, Chaillet M, Garzoni F, Berger I (2009) Getting a grip on complexes. Curr Genomics 10(8):558–572
14. Fernandez FJ, Vega MC (2013) Technologies to keep an eye on: alternative hosts for protein production in structural biology. Curr Opin Struct Biol 23(3):365–373

Part II

Prokaryotic Expression Hosts

3 ACEMBL Tool-Kits for High-Throughput Multigene Delivery and Expression in Prokaryotic and Eukaryotic Hosts

Yan Nie, Maxime Chaillet, Christian Becke, Matthias Haffke, Martin Pelosse, Daniel Fitzgerald, Ian Collinson, Christiane Schaffitzel, and Imre Berger

Abstract

Multicomponent biological systems perform a wide variety of functions and are crucially important for a broad range of critical health and disease states. A multitude of applications in contemporary molecular and synthetic biology rely on efficient, robust and flexible methods to assemble

Y. Nie
European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Unit of Virus Host-Cell Interactions, University Grenoble Alpes-EMBL-CNRS, UMI 3265, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Department of Structural Biochemistry, Max Planck Institute of Molecular Physiology, Otto-Hahn-Strasse 11, 44227 Dortmund, Germany

M. Chaillet • M. Pelosse
European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Unit of Virus Host-Cell Interactions, University Grenoble Alpes-EMBL-CNRS, UMI 3265, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

C. Becke
European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Department of Biochemistry, Freie Universität Berlin, Takustrasse 6, 14195 Berlin, Germany

M. Haffke
European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Unit of Virus Host-Cell Interactions, University Grenoble Alpes-EMBL-CNRS, UMI 3265, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Novartis Institutes for BioMedical Research, Novartis Campus, 4056 Basel, Switzerland

D. Fitzgerald
Geneva Biotech SARL, Avenue de la Roseraie 64, 1205 Genève, Switzerland

I. Collinson
School of Biochemistry, University of Bristol, Bristol BS8 1TD, UK

C. Schaffitzel • I. Berger (✉)
European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Unit of Virus Host-Cell Interactions, University Grenoble Alpes-EMBL-CNRS, UMI 3265, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

School of Biochemistry, University of Bristol, Bristol BS8 1TD, UK
e-mail: iberger@embl.fr

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_3

multicomponent DNA circuits as a prerequisite to recapitulate such biological systems *in vitro* and *in vivo*. Numerous functionalities need to be combined to allow for the controlled realization of information encoded in a defined DNA circuit. Much of biological function in cells is catalyzed by multiprotein machines typically made up of many subunits. Provision of these multiprotein complexes in the test-tube is a vital prerequisite to study their structure and function, to understand biology and to develop intervention strategies to correct malfunction in disease states. ACEMBL is a technology concept that specifically addresses the requirements of multicomponent DNA assembly into multigene constructs, for gene delivery and the production of multiprotein complexes in high-throughput. ACEMBL is applicable to prokaryotic and eukaryotic expression hosts, to accelerate basic and applied research and development. The ACEMBL concept, reagents, protocols and its potential are reviewed in this contribution.

Keywords

Gene delivery • High-throughput • Automation • Robotics • Structural proteomics • Protein complexes • Membrane proteins • Synthetic biology • Metabolic engineering

3.1 Complex Challenge: Functional Multigene Assembly and Delivery

Multigene delivery into living organisms has taken to center stage in the synthetic biology era [1, 2]. This development has been catalyzed by the emergence of powerful technologies to precisely assemble DNA pieces representing functional modules into customized multifunctional DNA circuits. Recombinant DNA technology emerged half a century ago, when so-called ‘restriction factors’ were observed, which inhibited bacteriophage growth in bacteria, which turned out to be DNA endonucleases [3–6]. Around this time, DNA ligation was discovered as a basis of genetic recombination [7–9], leading to successful assembly of DNA fragments [10–14]. Since these ground-breaking discoveries, classical DNA cloning involved largely serial steps of cutting and pasting isolated fragments together by using restriction enzymes and DNA ligases, into functional DNA molecules (typically plasmids). Plasmids containing the DNA insert of choice then were delivered by transformation, transduction or transfection into prokaryotic or eukaryotic host cell organisms to exert their functions [15]. The advent of the polymerase chain reaction (PCR) enormously advanced the field [16], making DNA cloning commonplace in virtually all molecular biology laboratories worldwide. Today, DNA assembly has been further accelerated by new and powerful technologies, including ligation independent cloning methods (LIC, SLIC) [17, 18], circular polymerase extension cloning (CPEC) [19] and seamless ligation cloning extract (SLiCE) [20], to name a few. For the assembly of very large fragments as precursors of entire synthetic genomes, specific cloning methods have been implemented [21]. Concomitantly, chemical DNA synthesis is being brought to perfection, considerably increasing the attainable size of DNA precursor fragments. These methods are at the core of synthetic biology, a vibrant field hailed as a game changer and poised to transform molecular biology and much of the life sciences [22–26].

Molecular cloning has been invaluable to study the structure and function of proteins by enabling heterologous expression. Elucidating the sequence content of entire genomes has made it possible to address the gene product repertoire—the proteome—of cells and organisms. Efficient DNA assembly methods to generate heterologous expression constructs have been implemented in concerted 'omics' efforts to analyze proteins system-wide, in high-throughput. Structural genomics consortia were established to determine atomic structures, seemingly in an industrial mode [27, 28]. Automation and robotics have become a prerogative; as a consequence, traditional cloning methodologies were progressively replaced by more advanced methods [1, 29–32]. A large number of vital functions in cells are mediated by multiprotein complexes composed of several to many subunits, and the function of a particular catalytic unit is often determined by its interaction partner(s). This has profound consequences for our understanding of the molecular mechanisms that are at the basis of biology. At the same time, this also imposes additional requirements on DNA assembly technology to support recombinant expression of complexes in high-throughput.

The ACEMBL technology was conceptualized to meet these requirements, to enable structural and functional "complexomics" research and discovery [31–34]. ACEMBL comprises recombination-based assembly of DNA elements into functional multigene expression constructs that can be rapidly permutated in a combinatorial fashion [33]. Originally, ACEMBL was developed for combinatorial multiprotein production in *E. coli* as a prokaryotic expression host [33]. Subsequently, efficient ACEMBL tool-kits have been developed also for multigene expression in eukaryotic hosts [34–37]. The integration of the ACEMBL technology in MultiBac, currently the lead technology for multiprotein complex production in insect cells, is described in a dedicated contribution of this issue [37]. The present overview therefore has as its focus the impact of ACEMBL on bacterial and mammalian multigene transfer applications.

3.2 ACEMBL: Automated Unrestricted DNA Recombineering for Multigene Delivery

Our knowledge of cellular processes has enormously advanced, brought about by an array of recent technological developments, notably in affinity purification, DNA sequencing, mass spectroscopy, yeast two-hybrid screens and computational approaches [38]. These technological developments compellingly validated the notion that virtually all essential cellular processes (DNA replication, transcription, translation, cell cycle regulation, intermediary metabolism, many more) are catalyzed by a highly coordinated network of protein-protein interactions, in which most proteins collaborate and function in the context of multiprotein complexes, underpinning the notion of 'protein sociology' in the cell [39].

Detailed structural and functional analysis is indispensable for elucidating the biological functions of these highly complex networks. Knowledge of molecular architectures can form the basis of intervention strategies, for example, to correct malfunction in disease states by supplying structure-based, custom-designed chemical compounds. Recapitulation of physiological interdependencies in the test-tube is a critical prerequisite for designing such compounds, and also for their preliminary validation by biochemical, biophysical and pharmacological means. Most multiprotein complexes, particularly those in humans, exist in (very) low endogenous amounts and are furthermore often heterogeneous in their composition, which is typically refractory to their extraction from native or cultured cell material. Heterogeneity in post-translational modifications, which may be essential for exerting the full activity of a given complex, can further limit the utility of material obtained from endogenous source.

Recombinant production offers solutions to these impediments, and a wide range of expression systems is available to produce proteins recombinantly in prokaryotic and eukaryotic hosts [31, 33–45]. Recombinant expression systems share in common that one or several DNA segments encoding for proteins, protein domains or multicomponent protein complexes are typically combined with DNA elements including DNAs that control transcription (promoters, terminators, others) and translation (ribosome binding sites, Shine-Dalgarno sequences, Kozak consensus sequences, enhancers, others) and inserted into a functional DNA module (plasmid, cosmid, artificial chromosome, genome, others). The resulting construct is then used to deliver the DNA segments of interest to host cell organisms by means of transformation, transfection or transduction.

The underlying technologies were perfected over the years to a level that they could be productively harnessed in ambitious, highly parallelized 'omics' programs aimed at genome- and proteome-wide studies of proteins in many organisms including humans [28]. In these research undertakings, protein encoding genes are synthesized, manipulated, varied and delivered into recombinant expression hosts on an industrial scale to enable high-throughput structure determination, populating protein structure databases such as the protein data bank (PDB) with unmatched efficiency and breathtaking speed, ushering in a new age of protein structural and functional research.

Initially, these efforts were focused on single proteins, protein domains or small assemblies of two, maximally (rarely) three interactors. Our more recent understanding that the activity of a given protein catalyst can be decisively influenced by the (sometimes many) partners that arrange in multicomponent assemblies has to a certain degree challenged this minimalist approach. It is legitimate to postulate that, if proteins in cells act as parts of large and complex assemblies, then they should also be studied *in vitro* in the form of such large complexes, with a full complement of binding partners present. This approach offers opportunities and advantages, notably for drug discovery in pharma and biotech, where 'being close(r) to physiological' can be a tremendous asset. Evidently, however, it also complicates the experimental approach quite significantly, posing substantial technical challenges.

A multigene delivery system that affords to establish physiologically meaningful contexts *ex vivo* needs to be simple to use, robust, efficient and ideally compatible with automation and robotics, and readily accessible if similar breakthroughs for multiprotein complexes are to be achieved as have been successfully made already for single proteins and protein domains. We have taken advantage of more recent advances in DNA synthesis and molecular cloning technologies to develop ACEMBL, a technology concept that in our view successfully addresses these challenges [33]. ACEMBL exploits sequence and ligation independent multifragment cloning technology combined with site-specific multicomponent recombination for unrestricted assembly of multigene delivery constructs in a combinatorial fashion that is readily amenable to robotics [33, 46]. Affordable and efficient chemical synthesis methods of large DNAs as precursor molecules further potentiate the utility of ACEMBL for a broad range of applications.

3.2.1 ACEMBL DNA Design

The ACEMBL system utilizes a series of custom-designed vectors (called Acceptor or Donor, respectively) for multigene vector generation catalyzed by Cre-LoxP recombination [33, 34, 44, 46, 47]. All ACEMBL vectors are scratch-built, synthetic small plasmids (2–3 kilobases). Acceptor and Donor plasmids exclusively contain the minimal DNA elements absolutely required for protein expression and plasmid propagation, in addition to a set of DNA elements required for multigene assembly. In contrast to conventional expression plasmids including most commercial plasmids, these elements are directly juxtaposed, without intervening sequences devoid of functionality, giving rise to the smallest possible DNA molecules that propagate and can be used for multigene expression (Fig. 3.1).

ACEMBL plasmids contain common modules such as promoter/terminator and resistance marker. The Multiple Integration Element (MIE), adapted from a previously published polylinker

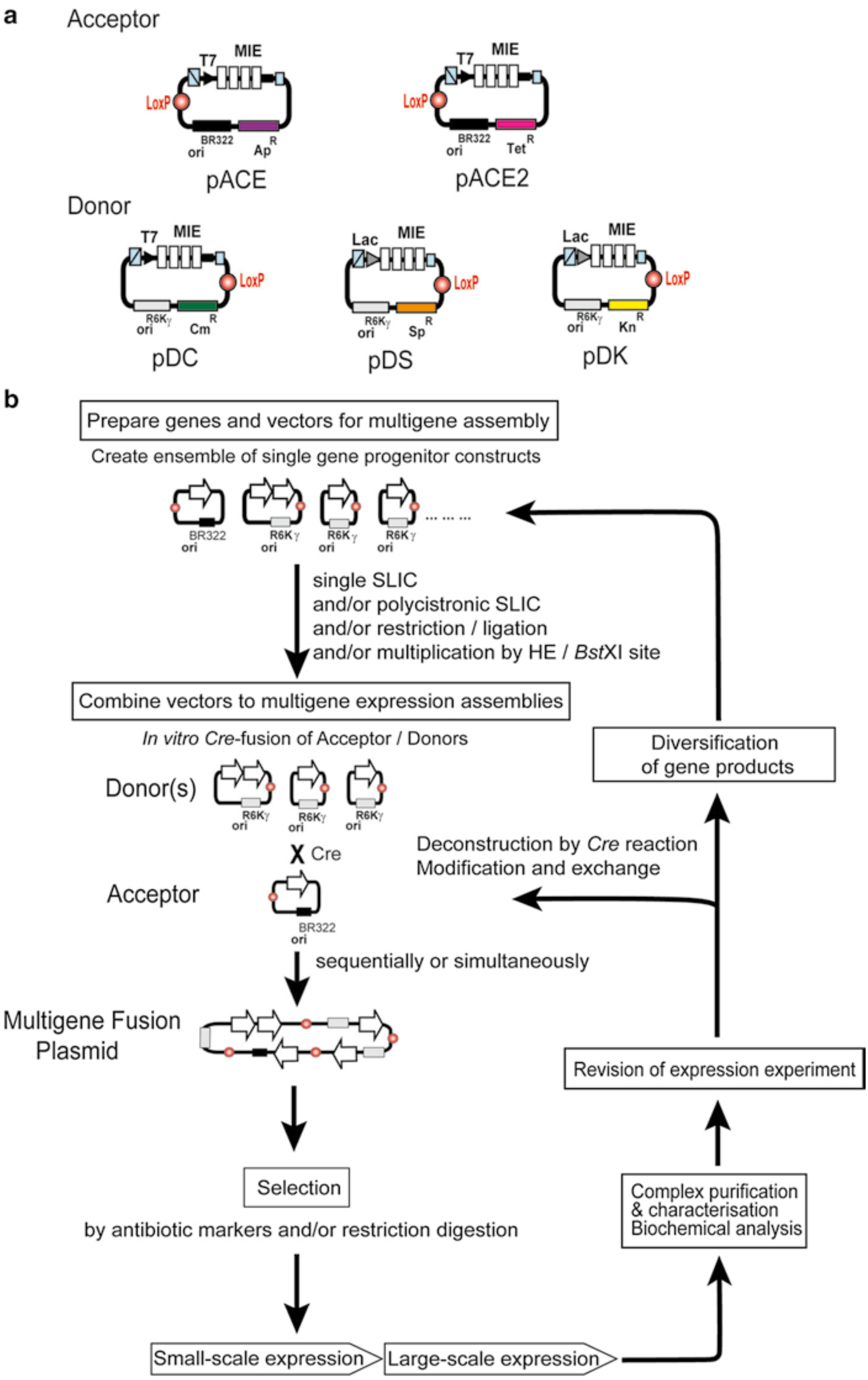

a
Acceptor
T7
MIE
LoxP
BR322
ori
Ap R
pACE
T7
MIE
LoxP
BR322
ori
Tet R
pACE2
Donor
T7
MIE
LoxP
R6Kγ
ori
Cm R
pDC
Lac
MIE
LoxP
R6Kγ
ori
Sp R
pDS
Lac
MIE
LoxP
R6Kγ
ori
Kn R
pDK
b
Prepare genes and vectors for multigene assembly
Create ensemble of single gene progenitor constructs
BR322 ori
R6Kγ ori
R6Kγ ori
R6Kγ ori
single SLIC
and/or polycistronic SLIC
and/or restriction / ligation
and/or multiplication by HE / BstXI site
Combine vectors to multigene expression assemblies
In vitro Cre-fusion of Acceptor / Donors
Donor(s)
R6Kγ ori
X Cre
Acceptor
BR322 ori
sequentially or simultaneously
Multigene Fusion Plasmid
Selection
by antibiotic markers and/or restriction digestion
Small-scale expression
Large-scale expression
Complex purification & characterisation Biochemical analysis
Revision of expression experiment
Diversification of gene products
Deconstruction by Cre reaction
Modification and exchange

Fig. 3.1 ACEMBL technology concept. (**a**) Acceptor and Donor plasmids are shown in a schematic view (*top*). The examples shown here are used for multigene delivery in *E. coli* as an expression host. Acceptor and Donor vectors contain a LoxP sequence and an identical Multiple Integration Element (MIE). Promoters (T7 or lac), corresponding terminators and homing endonuclease (HE) sites (*blue strike-through box*, Acceptors: I-CeuI; Donors: PI-SceI) and matching BstXI sites (*small blue squares*) are indicated. Origins of replication (Acceptors: BR322; Donors: R6Kγ) are shown. *Ap* Ampicillin, *Tet* Tetracycline, *Cm* Chloramphenicol, *Kn* Kanamycin, *Sp* Spectinomycin. The Multiple Integration Element (MIE) is specific for expression in a prokaryotic host and supports assembly of polycistrons encoding for several genes controlled by a single pair of promoter and terminator. (**b**) Outline of the method (Adapted from Ref. [33])

[33], is tailored to support single or multiple gene insertions via conventional restriction/ligation methods or, preferably, sequence and ligation independent cloning (SLIC) [33, 46] (Fig. 3.2). In addition, complementary homing endonuclease (HE)/BstXI site pairs are introduced for theoretically unlimited iterative gene insertions. We usually insert DNAs (genes of interest or fragments) that are chemically synthesized in the given format of choice, eliminating undesired restriction sites (including HE, BstXI) in the process. The expression cassettes in all ACEMBL plasmids are thus in a configuration which has been termed 'BioBrick' in synthetic biology applications, enabling multimerization.

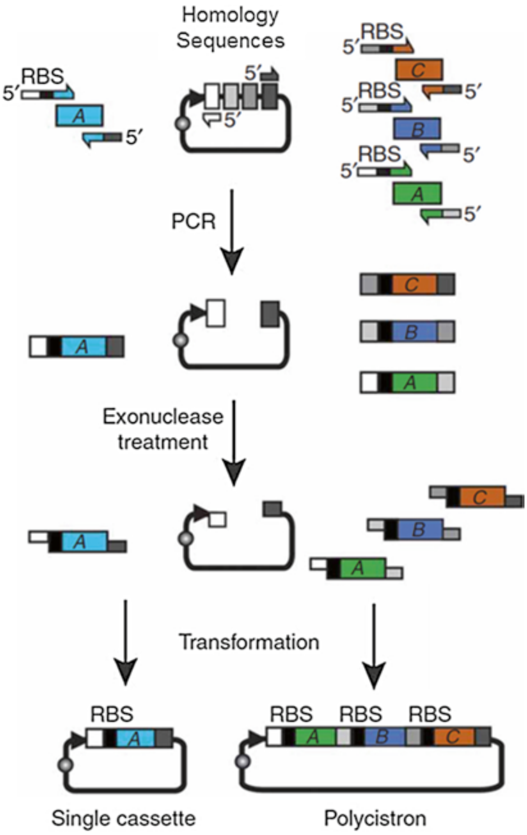


Fig. 3.2 Gene insertion into ACEMBL Acceptors and Donors by SLIC. Gene insertion into ACEMBL plasmids by sequence and ligation independent cloning (SLIC) is shown in a schematic representation. Primer DNA oligonucleotides used for PCR are shown as *thin bars* with *arrows*. RBS stands for ribosome binding site. 5′ denotes the five-prime end. Regions of homology in the Multiple Integration Element (MIE) are shown as boxes filled in *gray*. Single gene integration is shown on the *left*. Multigene integration yielding a polycistron is depicted on the *right*. PCR stands for polymerase chain reaction. Exonuclease treatment is conveniently performed by T4 DNA polymerase in the absence of deoxyribonucleotide triphosphates (dNTPs) (Adapted from Ref. [33])

There are two origins of replication in ACEMBL tool-kits; Acceptors contain a common *E. coli* origin of replication (BR322) and Donors contain a conditional origin of replication derived from phage R6Kγ. All plasmids contain a different resistance marker. Acceptors and Donors shown in Fig. 3.1 contain elements that are specific for multigene delivery and expression in *E. coli* as a prokaryotic host. Similar Acceptors and Donors have been developed for multigene delivery and multiprotein complex expression in eukaryotic hosts, retaining the backbones but containing customized DNA elements (promoter/terminator pairs, gene integration sites, homologous recombination sequences, others) required in the respective eukaryotic host organisms (mammalian and insect cells).

3.2.2 Multigene Assembly by Tandem Recombineering (TR)

The SLIC reaction, in marked contrast to conventional cloning relying on restriction enzyme mediated digestion and ligation, can be readily scripted into a robotics routine [33]. ACEMBL Acceptor and Donor plasmids that contain one or several genes each are then concatamerized for multigene co-expression in a rapid and flexible fashion, by utilizing the LoxP imperfect inverted repeat sequences present on each plasmid, and the Cre recombinase which fuses LoxP sequences in a site-specific recombination reaction (Fig. 3.3) [48, 49]. Tandem Recombineering (TR) is the combination of SLIC-mediated gene integration and Cre-LoxP Acceptor-Donor fusion [46].

When educt DNAs containing single LoxP sites are subjected to Cre-LoxP recombination, only a small portion of educt DNAs are combined together, while the rest remain separate and co-exist with the fusion products. Acceptors contain

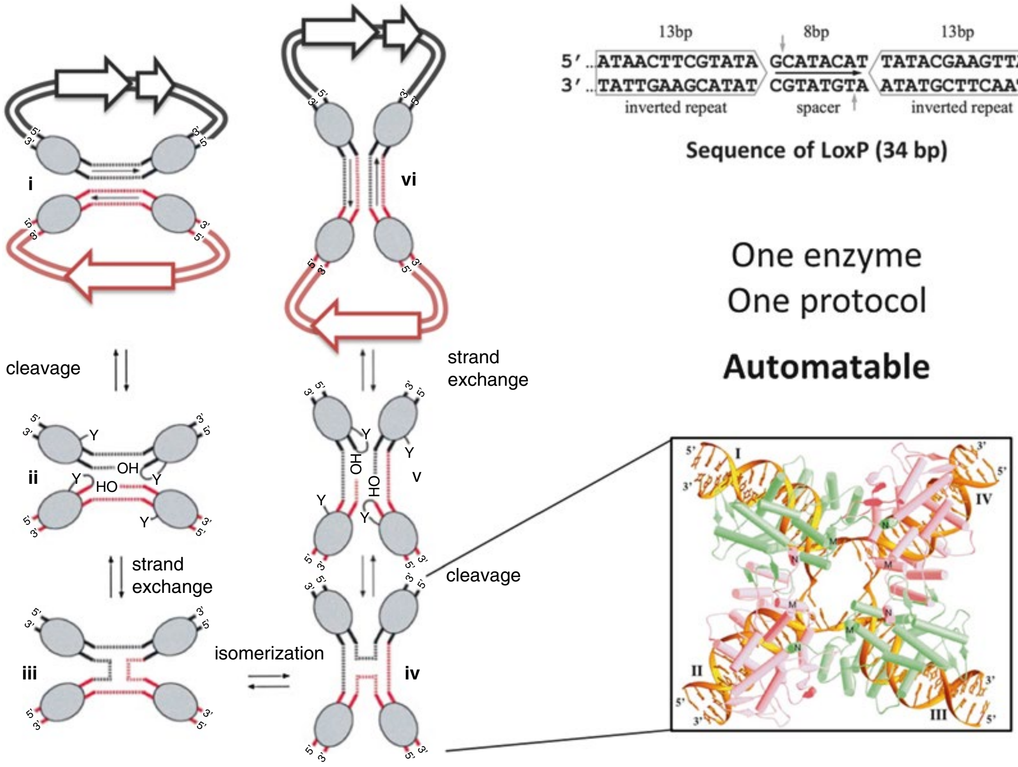


Fig. 3.3 Cre-LoxP fusion reaction. Mechanism of Cre-mediated DNA fusion is shown in a schematic representation. Cre enzyme (shape filled in *gray*) recognizes LoxP sites (marked by *dashed lines* and *arrow*) present on DNA molecules and fuses them in an equilibrium reaction favoring excision (*left*). The sequence of the LoxP imperfect inverted repeat is displayed (*top*, *right*). Cre-LoxP mediated fusion is a one-step reaction requiring a simple protocol that can be automated. The structure of four copies of Cre enzyme bound to a Holliday junction reaction intermediate is shown in the inset (Adapted from Ref. [49])

a regular origin of replication (BR322), which enables their replication in regular *E. coli* strains (TOP10, OmniMAX, BL21, etc.). In contrast, Donors contain a conditional origin of replication termed R6Kγ (the γ replication origin of the R6K plasmid) [50]. The replication of Donors requires the presence of the π protein (encoded by *pir* gene) in the host cell. Therefore, propagation and manipulation of all Donors has to be carried out in specific *E. coli* strains, which contain a *pir* gene inserted into their genome. Donors cannot replicate in a regular *E. coli* strain, which does not contain the *pir* gene (*i.e.*, *pir*-negative), unless fused with an Acceptor with a regular origin of replication. Thus, the recombination of Acceptors and Donors can be exploited for specific selection of desired fusion products.

A single Acceptor could be recombined in a single Cre-LoxP reaction with a theoretically unlimited number of Donors, with one to several expression cassettes on each Donor and Acceptor. Pragmatically, we use one Acceptor and up to three Donors to generate multigene constructs for heterologous expression. Due to the equilibrium nature of the Cre-LoxP reaction, the recombination reaction products are a mixture of all possible fusions from two or more educts, including Acceptor-Acceptor, Acceptor-Donor, and Donor-Donor fusions. Since excision is favored, fusion products containing increasing numbers of educts are present in decreasing amounts. All fusion products and also the single educt plasmids are quasi bar-coded by their characteristic resistance marker combinations (Fig. 3.4), as all plasmids

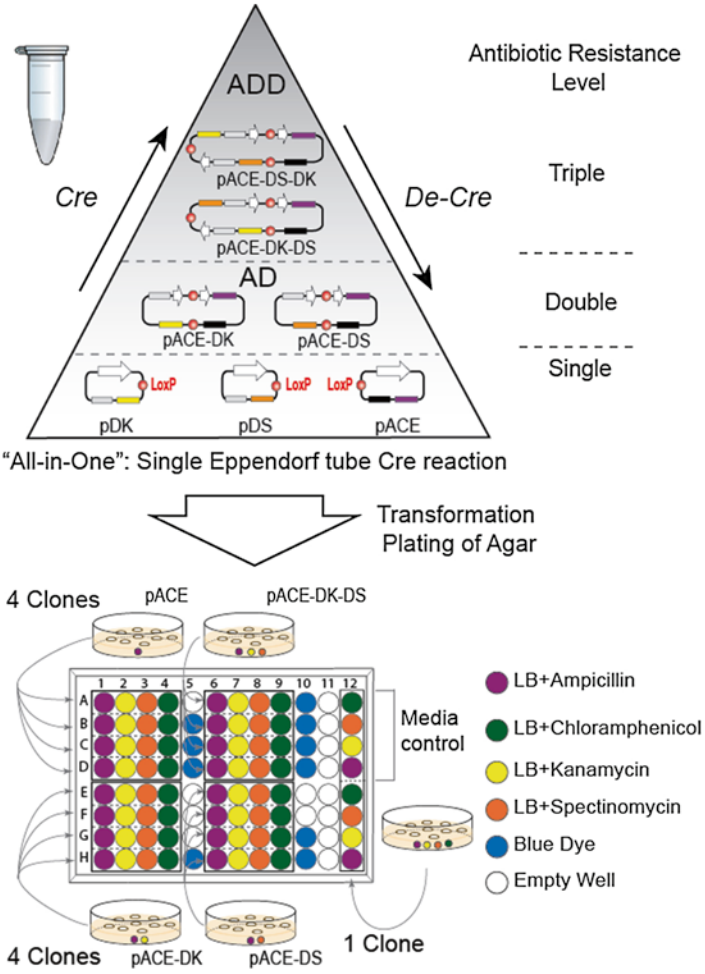


Fig. 3.4 ACEMBL combinatorics. Dynamic assembly (Cre) and disassembly (De-Cre) of Acceptor and Donor plasmids by Cre-LoxP reaction is shown schematically (*top*). LoxP sites are drawn as *red circles*; resistance markers and origins of replication are colored as above (Fig. 3.1). *White thick arrows* denote expression cassettes. AD stands for Acceptor-Donor fusion. ADD stands for Acceptor-Donor-Donor fusion. Not all possible fusion products are shown for clarity. Levels of multiresistance for product selection are indicated (*top*, *right*). All reactions occur in a single Eppendorf tube. Fusion products co-exist with educts. Productive fusion products are selected using (multi)antibiotic challenge, for example on a 96-well micro-titer plate (*bottom*). Desired Acceptor-Donor fusions are identified according to their resistance marker 'bar-code'. Color-coding of antibiotics is listed (*bottom*, *right*). LB stands for Luria-Bertani/lysogeny broth

of the system have a different resistance marker. After transformation into regular *E. coli* strains (*pir*-negative background), all unwanted Donors and Donor-Donor fusions are eliminated since their conditional origins are inactive in *pir*-negative *E. coli* strains, while the desired Acceptor-Donor fusions are selected by challenging with corresponding combinations of antibiotics (Fig. 3.4). This enables the tailored generation of multigene vectors expressing a complete protein complex as well as subsets of its subunits, in a single Cre-LoxP reaction. This combinatorial approach is instrumental for investigating the hierarchical assembly of multiprotein complexes, the biological functions of specific subunit(s) or their combinations, as well as the integration of putative subunit isoforms into a multiprotein complex of choice [31]. Thus selected arrays of fusion plasmids can then be used for gene delivery into expression host cells, optionally in high-throughput.

Subsequent to antibiotic challenge, fusion plasmids can (and probably should) be verified by restriction mapping. For example, transformants might contain fusion products harboring more than one copy of a particular educt vector. This can be potentially detrimental by causing expression level imbalance between subunits due to the increase in copy number of the gene(s) present on the particular educt. On the other hand, this could also be used to the benefit of the expression experiment. When a certain gene of

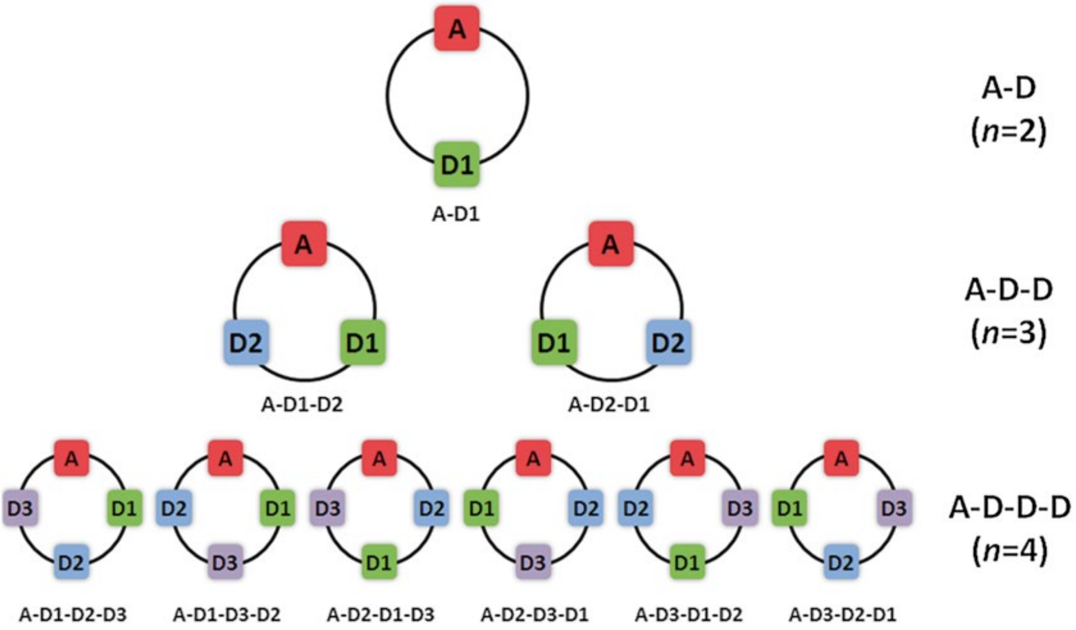


Fig. 3.5 Acceptor-Donor fusion arrays. Variants of possible multifusion plasmids are depicted, containing two (*top row*), three (*middle row*), or four (*bottom row*) educt plasmids (Acceptor, Donors), each as a single copy. Box filled in *red* denotes Acceptor (A), Boxes filled in *green*, *blue* and *purple* denote three Donors (D1, D2 and D3, respectively). The linear order (starting with A for simplicity) of educts in each (circular) multifusion plasmid is indicated below the corresponding plasmid map. The number of educt vectors and compositions are indicated (*right*)

interest is expressed at lower levels as compared to other genes in a multigene expression experiment, it can be helpful to incorporate an additional copy of the corresponding educt plasmid, and/or to place the same gene in several copies on one or more educt plasmids prior to the Cre-LoxP fusion reaction.

When more than two educt vectors are subjected to Cre-LoxP recombination, their incorporations are stochastic and thus lead to sequence variations in the fusion plasmids depending on the assembling orders of educt vectors (Fig. 3.5). The number of possible fusion plasmids (P_n) containing n educt vectors (each as a single copy) is given by the formula of circular permutation: $P_n = (n-1)!$. For example, a fusion plasmid containing one acceptor and three donors ($n=4$) has $P_4 = 3! = 6$ possible variants (Fig. 3.5). From our experience, the order of assembly of educts in a multifusion plasmid apparently does not prejudice the success of a complex expression experiment. Nonetheless, good practice requires verifying the order of assembly of educts in the multifusion plasmid as a quality control step. Therefore, the exact DNA sequences of all possible fusion variants are required for verification and selection by restriction digestions. To facilitate the *in silico* generation of DNA sequences of all possible fusion variants, we programmed a software application, Cre-ACEMBLER [51, 52].

3.2.3 Cre-ACEMBLER Software

Cre-ACEMBLER was programmed in Python and runs on Windows, Linux, and MacOS operating systems. Cre-ACEMBLER displays sequence data in an application window, showing the sequence as plain text. Simple manipulations can be done using cut, copy and paste functions. Sequence data can be read from and written to files in various formats, including FASTA and GenBank.

To perform *in silico* Cre recombinations, all educt plasmid sequences have to be opened in

Cre-ACEMBLER. Activating the “Cre” button starts an assistant dialogue guiding through the recombination in three steps: (1) Acceptor plasmid sequence is selected among all open sequences; (2) Donor plasmid sequences are selected; and (3) adjustment of the desired copy numbers of each individual plasmid. Each possible product sequence is then generated and displayed in a new window. Product sequences can then be saved to files and analyzed using other software, *e.g.*, ApE [53] or Vector NTI [54]. Prerequisites to be fulfilled by Cre-ACEMBLER were ease of use, compatibility with a broad range of operating systems and interoperability with other software. No central processing unit (CPU)-intensive work is done by Cre-ACEMBLER, thus an interpreted programming language could be chosen without risking performance limitations. Therefore, Cre-ACEMBLER was developed in Python [55], using the Python bindings of GTK+ [56, 57] for the graphical user interface, and the Biopython [58] library for sequence data manipulations. Using Python and GTK+ allows Cre-ACEMBLER to run on Windows, Linux and MacOS operating systems, and possibly others. The Biopython library allows reading and writing sequence data in various file formats, providing good interoperability with other software.

It is of advantage for the *in silico* Cre recombination if LoxP sites in all input (educt) sequences are in the same orientation, and if the linear representation of each input sequence starts with the LoxP site. Therefore, all input sequences are normalized prior to recombination, by generating the reverse-complement of input sequences if required, and by linearizing all sequences immediately 5′ of the LoxP site. All input sequences are then indexed numerically, making sure that identical input sequences get the same index. Lists representing all possible permutations of the order of the indices are computed, and redundant solutions (if considering circular arrangement) are eliminated, thus yielding index lists representing only unique circular permutations. Fusion plasmid sequences are then generated from these index lists by appending the normalized input sequences corresponding to the indices, in the order given in these lists.

A challenge arising from the linear representation of circular sequences is to identify permutations which are redundant if circular arrangement is considered. In order to make the lists representing different circular arrangements comparable, a linearization algorithm had to be found which transforms a linear representation with a random starting point reliably into a linear representation with a defined starting point. To accomplish this, the lowest index in the lists is taken as a potential starting point for linearization. If several instances of this lowest index are present in the list, each instance is credited a score according to the subsequent indices in the list. The instance that is followed by the highest count of lowest indices gets the highest score, and the list is rearranged such that this instance becomes the first entry. Lists transformed in this way can then simply be compared using Python’s equality operator, so that redundant solutions can be identified and eliminated.

Cre-ACEMBLER has proven to be a valuable, robust tool in extensive testing by users of the Eukaryotic Expression Facility (EEF) at EMBL Grenoble, proving the reliability of the algorithms described above. Cre-ACEMBLER is freely available for download [51]. A Cre-ACEMBLER User Manual is likewise available on-line [52].

3.3 ACEMBL Applications

The ACEMBL system was first introduced for robotized production of multiprotein complexes in high-throughput [33]. Subsequently, the ACEMBL pipeline was extended to eukaryotic expression systems (Fig. 3.6) in order to produce functional eukaryotic protein complexes requiring the authentic processing and post-translational machinery provided by eukaryotic hosts [31]. Multifusion plasmids generated from Cre-LoxP reactions are utilized by the ACEMBL-derived MultiMam system to facilitate simultaneous multigene introduction into mammalian cells [35, 36] (see also Sect. 3.2). The MultiBac baculovirus/insect cell system has been upgraded for robotics by incorporating ACEMBL DNA

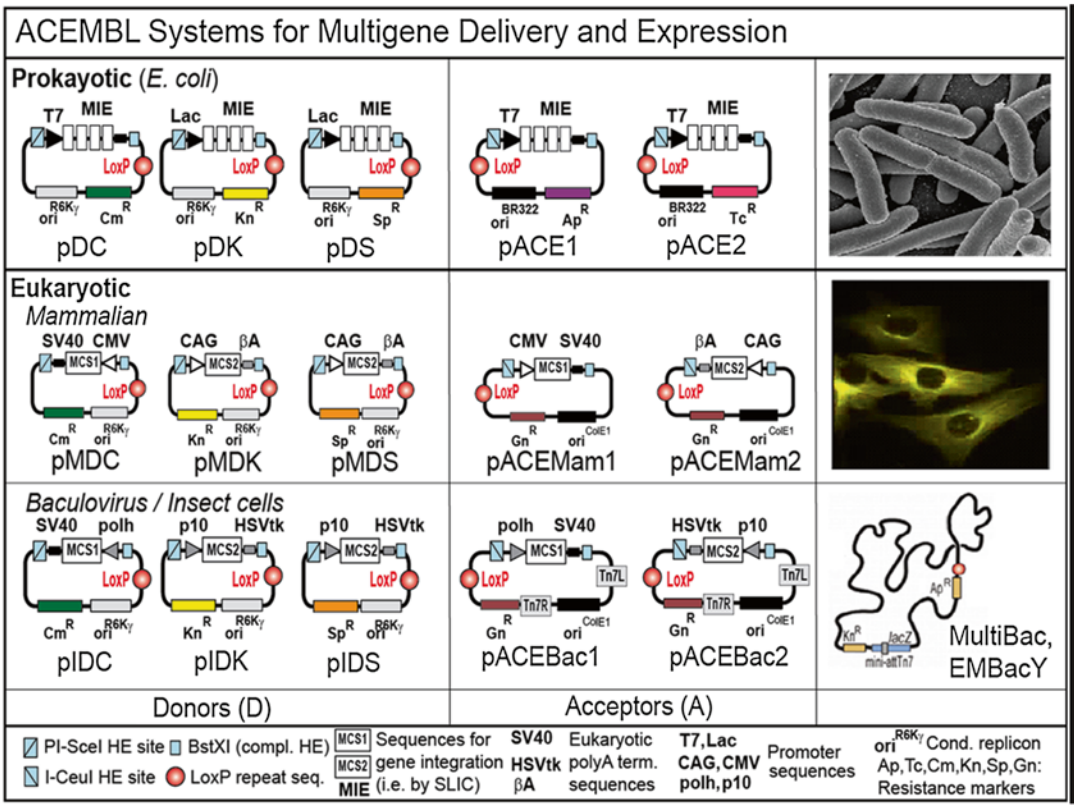


Fig. 3.6 ACEMBL tool-kits (as of 2014). Prokaryotic and eukaryotic expression systems derived from ACEMBL technology for multiprotein co-expression (Adapted from Ref. [31]). Note that initially, ACEMBL referred to the *E. coli* system. We have now named the individual ACEMBL systems MultiColi for *E. coli*, MultiMam for mammalian and MultiBac for baculovirus/insect cell expression. Expression cassettes in all ACEMBL plasmids were functionalized in 'BioBrick' format, enabling iterative multiplication

modules (MIE and HE/BstXI sites) for automatable and theoretically unlimited multigene insertion into a baculoviral genome for protein co-expression in insect cells [34] (see also Chap. 13 on MultiBac by Sari and co-workers). Selected examples of ACEMBL applications are highlighted in the following.

3.3.1 ACEMBLing DNA for Structural and Molecular Biology

ACEMBL has been used successfully for a variety of applications in structural and molecular biology. Numerous multisubunit complexes, including soluble multiprotein complexes, protein-RNA complexes and multimeric membrane protein complexes have been produced successfully by ACEMBL [33, 44, 59–63]. Examples include the prokaryotic signal recognition particle, SRP, the catalytic cycle of which is being studied by cryo-electron microscopy and biochemical means [59–61]. A particular highlight is the prokaryotic holo-translocon complex (HTL), a seven subunit transmembrane multiprotein assembly consisting of the heterotrimeric core translocon, SecYEG, and its accessory proteins SecD, SecF, YidC and YajC. HTL is a long elusive complex that was, for the first time, successfully produced recombinantly by ACEMBL [62, 63]. HTL catalyzes the transport of protein substrates through and into membranes, making use of the proton motive force (PMF) [62]. Moreover, ACEMBL was applied to reveal the substrate specificity for interferon-stimulated gene 15 by ubiquitin-specific protease 18 [64]. Many research laboratories have already obtained ACEMBL reagents, and ACEMBL systems are in the process of being integrated into structural genomics pipelines. We expect in the coming years numerous more exploits brought about by our multigene delivery technologies, and we anticipate productive synergies with other multigene recombineering tools, to deconvolute internal redundancy and explore functional structure in complex biological systems [65–68].

3.3.2 Highly Efficient Multigene Delivery in Mammalian Cells

We implemented TR to facilitate rapid generation of multicomponent gene expression circuits from Acceptors and Donors containing mammalian cell active promoters [35, 36] (Fig. 3.6). These multicomponent circuits are used for efficient multigene delivery in mammalian cells, resulting in homogeneous cell populations [35]. Such results could not be obtained previously by classical methods relying on co-transfection of plasmids modules. Using fluorescently labeled proteins to visualize mammalian cell compartments, their substructures and contents is a common technology in cell biology and pharmacological applications. Homogeneous cell populations are a prerequisite for monitoring perturbations of cell states, biological processes, metabolic pathways, signaling cascades and the effect of additives, for example in high-content screening. The utility of the TR approach to generate homogeneous cell populations by multigene delivery of fluorescently labeled proteins was compellingly demonstrated using pig cardiac endothelial cells that expressed five different proteins, delivered by a TR construct fitted with mammalian cell active promoters [35]. A constant relationship between expression levels of the proteins at the level of individual cells was demonstrated [35]. Moreover, this approach was applied to analyze the localization of epidermal growth factor receptor (EGFR) with Ran GTPases in endosomal trafficking, and to demonstrate how Neuropilin-1 promotes VEGFR-2 trafficking through Rab11 vesicles, thereby specifying signal output [35, 69].

We anticipate that a wide range of applications will benefit from a synchronized delivery of multiple genes. Our innovative approach has the potential to facilitate the production of multicomponent protein drugs including next-generation vaccine candidates such as virus-like particles. Multiplexed labeling of living cells, protein-protein interaction studies, the construction of designed gene regulatory circuits and entire synthetic signaling cascades are further active research and development fields that could benefit from ACEMBL technologies.

3.4 Metabolic Engineering

Metabolic engineering is emerging as an overarching concept subsuming a collection of methods and concepts for re-directing, improving or modifying cellular and organismal biochemical pathways, with the goal of generating novel qualities. At the core of synthetic biology, metabolic engineering has been defined as “the purposeful modification of cellular activities with the aim of strain improvement” [70]. A purpose is to achieve an overall higher productivity and superior quality of scientifically or commercially interesting molecules in research and development, and in industrial settings. These can include proteins, protein complexes, nucleic acids, biochemicals and metabolites that normally do not accumulate to a significant degree or sufficient quality, and would otherwise have to be chemically synthesized or extracted from natural sources. Moreover, complex chemical structures, for instance with multiple chiral centers, often are more easily produced in engineered microorganisms or cells, at lower cost.

Researchers wish to tweak the host which is the organismal “factory”, by altering its biological traits, to produce modified or new substances [71]. Such refinements require considerable genetic engineering for custom-design of entire regulatory circuits and metabolic pathways and their efficient delivery into the host organism. Concurrently, ‘negative’ factors need to be removed, which would otherwise be detrimental to achieving the desired product yields and quality improvements. For example, production strains may need to be made more resilient to demands incurred by (multi)protein overexpression [72]. Physiological knowledge of the pathways under investigation, choice of the right production organism, information from metabolic flux modeling and bioprocess development need to be considered and addressed in the design of the synthetic multifunctional DNA circuits to be delivered [73]. This can then be exploited for the improved production, up to fermenter scale, of protein therapeutics such as monoclonal antibodies, commodity chemicals such as vitamins or rare amino acids, valuable metabolites, biomolecules eliciting fragrances and flavors, rare natural (medicinal) compounds (such as artemisinin and taxol) or even biofuel production [73–77]. While some constraints can be overcome by optimizing culture conditions, others can more successfully be tackled by modifying defined metabolic pathways. A thorough knowledge of the cellular biochemistry in conjunction with new and powerful recombinant DNA technology now allows “to rationally modify and design metabolic pathways, proteins, and even whole organisms.” [77].

Biosynthetic pathways can now be (re)constructed from scratch and adapted to a host organism to either replace or complement endogenous pathways [74, 78]. Genetic modifications involved include for example plugging in appropriate regulatory elements into the plasmid constructs, optimizing codon usage or transcription factor activity, and tuning the effects of intergenic regions. At the same time, endogenous pathways can be shut down or reduced [79] to optimize the balance between heterologous and endogenous biochemical activities [78]. Side effects or roadblocks encountered can be ameliorated or removed by multiple rounds of engineering [78, 80, 81].

ACEMBL tool-kits, due to unmatched flexibility and robustness, in our view may be optimally suited to address these manifold requirements for building multifunctional heterologous expression constructs, predominantly to equip *E. coli*, insect and mammalian cells with multiple genes and functionalities, combinatorially arranged by TR in multicomponent DNA regulatory circuits. An advantage of ACEMBL is that individual (sets of) components can be distributed on several plasmid modules (Acceptors and multiple Donors) and recombined as desired. Furthermore, individual (sets of) components can be flexibly modified without compromising other (sets of) components, and new components introduced if required. Moreover, gene regulatory elements including promoters and terminators can be altered or tuned with ease, and adapted to the host organism and the specific requirements of the target molecule(s).

3.5 Conclusion

The ACEMBL technology concept was originally conceived to synergistically address two sets of requirements. On the one hand, we intended to create technologies that assist in making hitherto inaccessible target molecules, in particular multiprotein complexes, amenable to high-resolution structural and functional analysis as a prerequisite to better understand their cellular activities, and to enable their modulation for example if malfunction occurs in disease states. On the other hand, we wanted our technologies to be sufficiently robust to facilitate automation and robotics, to harness the benefits of parallelized workflows that already have been established for high-throughput applications with remarkable success. ACEMBL fulfills these requirements, and we are hopeful that the methods we developed will contribute significantly to the system-wide elucidation of the protein ‘complexome’ of cells and organisms, in health and disease. Currently, ACEMBL reagents are available for multigene delivery and heterologous expression in *E. coli*, mammalian cells and insect cells as hosts, and further systems targeting other important organismal factories are forthcoming. Moreover, beyond heterologous protein complex production, ACEMBL holds significant promise to catalyze synthetic biology approaches which are at the forefront of current biology, by enabling multiplexed assembly of synthetic multicomponent DNA constructions for highly efficient multigene delivery, in prokaryotic and eukaryotic hosts, for a wide range of applications.

Acknowledgments We thank all members of the Berger laboratory for their support, as well as Lakshmi S. Vijayachandran (Amrita Center for Nanosciences & Molecular Medicine, India), Christoph Bieniossek (Hofmann-La Roche, Basel), Simon Trowitzsch (Goethe University, Frankfurt) and Alexander Craig (S. Karger GmbH, Publishers) for their contribution and helpful discussions. This work was funded by the European Commission (EC) Framework Programme (FP) 7 ComplexINC project (grant number 279039) to DF and IB. YN was a fellow of the Boehringer Ingelheim Fonds (BIF, Germany). MH was recipient of a Kekulé fellowship from the Fonds der Chemischen Industrie (FCI, Germany).

References

1. Schelshorn D, Ljubicic S, Berger I, Fitzgerald DJ (2015) Synthetic biology in biopharmaceutical production. Eur Biopharm Rev 12(228):3995–3997
2. Tirabassi R (2014) Foundations of molecular cloning – past, present and future. http://www.neb-online.fr/pdfs/Article_MolecularCloning2014_NEBFR.pdf
3. Luria SE, Human ML (1952) A nonhereditary, host-induced variation of bacterial viruses. J Bacteriol 64:557–569
4. Bartani G, Weigle JJ (1953) Host controlled variation in bacterial viruses. J Bacteriol 65:113–121
5. Linn S, Arber W (1968) Host specificity of DNA produced by Escherichia coli, X. In vitro restriction of phage fd replicative form. Proc Natl Acad Sci U S A 59:1300–1306
6. Smith HO, Wilcox KW (1970) A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J Mol Biol 51:379–391
7. Kellenberger G, Zichichi ML, Weigle JJ (1961) Exchange of DNA in the recombination of bacteriophage lambda. Proc Natl Acad Sci U S A 47:869–878
8. Meselson M, Weigle JJ (1961) Chromosome brekage accompanying genetic recombination in bacteriophage. Proc Natl Acad Sci U S A 47:857–868
9. Bode VC, Kaiser AD (1965) Changes in the structure and activity of lambda DNA in a superinfected immune bacterium. J Mol Biol 14:399–417
10. Cozzarelli NR, Melechen NE, Jovin TM, Kornberg A (1967) Polynucleotide cellulose as a substrate for a polynucleotide ligase induced by phage T4. Biochem Biophys Res Commun 28:578–586
11. Gefter ML, Becker A, Hurwitz J (1967) The enzymatic repair of DNA. I. Formation of circular lambda-DNA. Proc Natl Acad Sci U S A 58:240–247
12. Gellert M (1967) Formation of covalent circles of lambda DNA by E. coli extracts. Proc Natl Acad Sci U S A 57:148–155
13. Olivera BM, Lehman IR (1967) Linkage of polynucleotides through phosphodiester bonds by an enzyme from Escherichia coli. Proc Natl Acad Sci U S A 57:1426–1433
14. Weiss B, Richardson CC (1967) Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage. Proc Natl Acad Sci U S A 57:1021–1028
15. Maniatis T (2012) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, New York. ISBN 978-1-936113-42-2
16. Mullis KB, Faloona FA (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 155:335–350
17. Li MZ, Elledge SJ (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nat Methods 4:251–256
18. Nisson PE, Rashtchian A, Watkins PC (1991) Rapid and efficient cloning of Alu-PCR products using uracil DNA glycosylase. PCR Methods Appl 1:120–123
19. Quan J, Tian J (2009) Circular polymerase extension cloning of complex gene libraries and pathways. PLoS One 4(7):e6441
20. Zhang Y, Werling U, Edelmann W (2012) SLiCE: a novel bacterial cell extract-based DNA cloning method. Nucleic Acids Res 40(8):e55
21. Gibson DG (2014) Programming biological operating systems: genome design, assembly and activation. Nat Methods 11(5):521–526
22. Weber W, Fussenegger M (2011) Emerging biomedical applications of synthetic biology. Nat Rev Genet 13(1):21–35
23. Ye H, Aubel D, Fussenegger M (2013) Synthetic mammalian gene circuits for biomedical applications. Curr Opin Chem Biol 17(6):910–917
24. Folcher M, Fussenegger M (2012) Synthetic biology advancing clinical applications. Curr Opin Chem Biol 16(3–4):345–354
25. Ausländer S, Fussenegger M (2013) From gene switches to mammalian designer cells: present and future prospects. Trends Biotechnol 31(3):155–168
26. Geering B, Fussenegger M (2015) Synthetic immunology: modulating the human immune system. Trends Biotechnol 33(2):65–79
27. Terwilliger TC, Stuart D, Yokoyama S (2009) Lessons from structural genomics. Annu Rev Biophys 38:371–383
28. Almo SC et al (2013) Protein production from the structural genomics perspective: achievements and future needs. Curr Opin Struct Biol 23(3):335–344
29. Torella J et al (2014) Unique nucleotide sequence-guided assembly of repetitive DNA parts for synthetic biology applications. Nat Protoc 9:2075–2089

30. Leinert F et al (2013) Two- and three-input TALE-based and logic computation in embryonic stem cells. Nucleic Acids Res 41:9967–9975
31. Vijayachandran LS et al (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J Struct Biol 175(2):198–208
32. Trowitzsch S, Palmberger D, Fitzgerald D, Takagi Y, Berger I (2012) MultiBac complexomics. Expert Rev Proteomics 9(4):363–373
33. Bieniossek C et al (2009) Automated unrestricted multigene recombineering for multiprotein complex production. Nat Methods 6(6):447–450
34. Bieniossek C, Imasaki T, Takagi Y, Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 37(2):49–57
35. Kriz A et al (2011) A plasmid-based multigene expression system for mammalian cells. Nat Commun 1:e120
36. Trowitzsch S, Klumpp M, Thoma R, Carralot JP, Berger I (2011) Light it up: highly efficient multigene delivery in mammalian cells. Bioessays 33(12):946–955
37. Sari D et al (2015) The MultiBac baculovirus/insect cell expression vector system for producing complex protein biologics. In: Vega C, Fernandez F (eds) Advances in experimental medicine and biology. Springer, New York
38. Nie Y et al (2009) Getting a grip on complexes. Curr Genomics 10(8):558–572
39. Robinson CV, Sali A, Baumeister W (2007) The molecular sociology of the cell. Nature 450(7172):973–982
40. Romier C et al (2006) Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. Acta Crystallogr D Biol Crystallogr 62(10):1232–1242
41. Busso D et al (2011) Expression of protein complexes using multiple Escherichia coli protein co-expression systems: a benchmarking study. J Struct Biol 175(2):159–170
42. Diebold ML et al (2011) Deciphering correct strategies for multiprotein complex assembly by co-expression: application to complexes as large as the histone octamer. J Struct Biol 175(2):178–188
43. Vincentelli R, Romier C (2013) Expression in Escherichia coli: becoming faster and more complex. Curr Opin Struct Biol 23(3):326–334
44. Haffke M et al (2015) Characterization and production of protein complexes by co-expression in Escherichia coli. Methods Mol Biol 1261:63–89
45. Abdulrahman W et al (2015) The production of multiprotein complexes in insect cells using the baculovirus expression system. Methods Mol Biol 1261:91–114
46. Haffke M, Viola C, Nie Y, Berger I (2013) Tandem recombineering by SLIC cloning and Cre-LoxP fusion to generate multigene expression constructs for protein complex research. Methods Mol Biol 1073:131–140
47. Berger I (2010) New nucleic acid tools for producing multiprotein complexes. WO2010/100278-A2
48. Guo F, Gopaul DN, van Duyne GD (1997) Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature 389(6646):40–406
49. Gopaul DN, Guo F, Van Duyne GD (1998) Structure of the Holliday junction intermediate in Cre-loxP site-specific recombination. EMBO J 17(14):4175–4187
50. Metcalf WW, Jiang W, Wanner BL (1994) Use of the rep technique for allele replacement to construct new Escherichia coli hosts for maintenance of R6K gamma origin plasmids at different copy numbers. Gene 138(1–2):1–7
51. Cre-ACEMBLR software: https://github.com/christianbecke/Cre-ACEMBLER
52. Becke C, Haffke M, Berger I (2012) Cre-ACEMBLER User Manual. doi:10.13140/2.1.1068.1128
53. Davis MW. ApE – a plasmid editor. http://www.biology.utah.edu/jorgensen/wayned/ape
54. Life Technologies Invitrogen Vector NTI. http://www.invitrogen.com
55. Python programming language. http://python.org/
56. The GIMP Toolkit. http://www.gtk.org/
57. PyGTK: GTK+ for Python. http://www.pygtk.org/
58. Cock PJA et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423
59. Schaffitzel C et al (2006) Structure of the *E. coli* signal recognition particle bound to a translating ribosome. Nature 444:503–506
60. Estrozi LF, Boehringer D, Shan SO, Ban N, Schaffitzel C (2011) Cryo-EM structure of the E. coli translating ribosome in complex with SRP and its receptor. Nat Struct Mol Biol 18(1):88–90
61. Von Loeffelholz O et al (2013) Structural basis of signal sequence surveillance and selection by the SRP-SR complex. Nat Struct Mol Biol 20:604–610
62. Schulze RJ et al (2014) Membrane protein insertion and proton-motive-force-dependent secretion through the bacterial holo-translocon SecYEG-SecDF-YajC-YidC. Proc Natl Acad Sci U S A 111(13):4844–4849
63. Komar J, Botte M, Collinson C, Schaffitzel C, Berger I (2015) ACEMBLing a multiprotein transmembrane complex: the functional SecYEG-SecDFYajC-YidC holotranslocon protein secretase/insertase. Methods Enzymol 556:23–49
64. Basters A et al (2014) Molecular characterization of ubiquitin-specific protease 18 reveals substrate specificity for interferon-stimulated gene 15. FEBS J 281(7):1918–1928
65. Cunna S et al (2011) Genetic disassembly and combinatorial reassembly identify a minimal functional repertoire of type III effectors in Pseudomonas syringae. Proc Natl Acad Sci U S A 108(7):2975–2980
66. Jakobi AJ, Huizinga EG (2012) A rapid cloning method employing orthogonal end protection. PLoS One 7(6):e37617

67. Zheng N, Huang X, Yin B, Wang D, Xie Q (2012) An effective system for detecting protein-protein interaction based on in vivo cleavage by PPV NIa protease. Protein Cell 3(12):921–928
68. Rode AB, Endoh T, Sugimoto N (2015) Tuning riboswitch-mediated gene regulation by rational control of aptamer ligand binding properties. Angew Chem Int Ed Eng 54(3):905–909
69. Ballmer-Hofer K, Andersson AE, Ratcliffe LE, Berger P (2011) Neuropilin-1 promotes VEGFR-2 trafficking through Rab11 vesicles thereby specifying signal output. Blood 118(3):816–826
70. Gosset G (2005) Improvement of Escherichia coli production strains by modification of the phosphoenolpyruvate:sugar phosphotransferase system. Microb Cell Factories 4:14
71. Patnaik R (2008) Engineering complex phenotypes in industrial strains. Biotechnol Prog 24:38–47
72. Chou CP (2007) Engineering cell physiology to enhance recombinant protein production in *Escherichia coli*. Appl Microbiol Biotechnol 76:521–532
73. Lee SY, Kim HU, Park JH, Park JM, Kim TY (2009) Metabolic engineering of microorganisms: general strategies and drug production. Drug Discov Today 14:78–88
74. Chang MCY, Keasling JD (2006) Production of isoprenoid pharmaceuticals by engineered microbes. Nat Chem Biol 2:674–681
75. Chemler JA, Koffas MAG (2008) Metabolic engineering for plant natural product biosynthesis in microbes. Curr Opin Biotechnol 19:597–605
76. Lee SK, Chou H, Ham TS, Lee TS, Keasling JD (2008) Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. Curr Opin Biotechnol 19:556–563
77. Jarboe LR et al (2010) Metabolic engineering for production of biorenewable fuels and chemicals: contributions of synthetic biology. J Biomed Biotechnol 2010:761042–761060
78. Pitera DJ, Paddon CJ, Newman JD, Keasling JD (2007) Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. Metab Eng 9:193–207
79. Alper H, Miyaoku K, Stephanopoulos G (2005) Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets. Nat Biotechnol 23:612–616
80. Berríos-Rivera SJ, Bennett GN, San KY (2002) Metabolic engineering of *Escherichia coli*: increase of NADH availability by overexpressing an NAD$^+$-dependent formate dehydrogenase. Metab Eng 4:217–229
81. Park JH, Lee KH, Kim TY, Lee SY (2007) Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. Proc Natl Acad Sci U S A 104:7797–7802

Complex Reconstitution and Characterization by Combining Co-expression Techniques in *Escherichia coli* with High-Throughput

4

Renaud Vincentelli and Christophe Romier

Abstract

Single protein expression technologies have strongly benefited from the Structural Genomics initiatives that have introduced parallelization at the laboratory level. Specifically, the developments made in the wake of these initiatives have revitalized the use of *Escherichia coli* as major host for heterologous protein expression. In parallel to these improvements for single expression, technologies for complex reconstitution by co-expression in *E. coli* have been developed. Assessments of these co-expression technologies have highlighted the need for combinatorial experiments requiring automated protocols. These requirements can be fulfilled by adapting the high-throughput approaches that have been developed for single expression to the co-expression technologies. Yet, challenges are laying ahead that further need to be addressed and that are only starting to be taken into account in the case of single expression. These notably include the biophysical characterization of the samples at the small-scale level. Specifically, these approaches aim at discriminating the samples at an early stage of their production based on various biophysical criteria leading to cost-effectiveness and time-saving. This chapter addresses these various issues to provide the reader with a broad and comprehensive overview of complex reconstitution and characterization by co-expression in *E. coli*.

R. Vincentelli (✉)
Architecture et Fonction des Macromolécules
Biologiques (A.F.M.B), UMR7257 CNRS,
Université Aix-Marseille,
Case 932, 163 Avenue de Luminy,
13288 Marseille cedex 9, France
e-mail: renaud.vincentelli@afmb.univ-mrs.fr

C. Romier (✉)
Département de Biologie Structurale Intégrative,
Centre de Biologie Intégrative, Institut de Génétique
et Biologie Moléculaire et Cellulaire (IGBMC),
Université de Strasbourg (UDS), CNRS, INSERM,
1 rue Laurent Fries, B.P. 10142, 67404 Illkirch
Cedex, France
e-mail: romier@igbmc.fr

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_4

Keywords

Protein complexes • Co-expression • *Escherichia coli* • Expression vectors • Expression parameters • Purification • Automation • Biophysics • Structural characterization

4.1 Introduction

Most of the functional units within cells are formed by macromolecular complexes. The intricate architectural and functional nature of these complexes highlights the challenge faced by researchers in studying these functional players. One of the major bottlenecks of these studies is to reconstitute and purify these complexes as homogeneous samples amenable to biochemical, biophysical and structural characterizations. Of particular importance is the requirement to obtain complexes that display low heterogeneity in terms of composition, structure and function.

Different technologies have been developed to help reconstitute and purify macromolecular complexes. These range from the purification of endogenous complexes to the overexpression of their subunits in endogenous or heterologous hosts. In the case of overexpression, assembly of the complexes can be done either in vivo, through co-expression techniques [1–5], or in vitro, by mixing independently purified subunits (see Chap. 19 by Jérôme Basquin in this issue).

These different technologies all have their own advantages and drawbacks. The choice for one technology rather than another will be driven by the project conducted, but also by the facility in implementing the technique to be used. Specifically, the co-expression technology has received particular attention in the last decades since it combines the advantage of in vivo complex assembly with the possibility, like for in vitro reconstitution approaches, of considering only a subset of the complex subunits, either full-length or truncated, to help obtain well-behaving homogeneous samples.

Escherichia coli remains a major host for protein production in most laboratories. This focus comes from the ease of *E. coli* genetic modification and its inexpensive and flexible use as expression host. In addition, the use of *Escherichia coli* has been revitalized by the development of Structural Genomics initiatives that have introduced automation and parallelization at the laboratory level, and have developed small-scale effective handling protocols well adapted to *E. coli* [6–9]. A similar trend is observed for co-expression technologies using *E. coli* as expression host [2, 4, 10–22]. Importantly, assessments of the co-expression technologies have highlighted the need for combinatorial experiments requiring automated protocols [1–5]. This requirement should be fulfilled by adapting the high-throughput approaches developed for *E. coli* single expression to the co-expression technologies [4].

Yet, the use of automated approaches at the small scale level often provides a large set of solutions in terms of construct to be used and in expression (*e.g.*, media, temperature, helper plasmids [23]) and purification (*e.g.*, pH, salt) parameters. The challenge is then to choose the few solutions that will be the most useful for successfully pursuing the project. If this issue is important for single expression, it is even more central when considering complexes and co-expression. The combinatorial nature of complex reconstitution by co-expression, *i.e.* considering various proteins and several constructs per protein, can very quickly yield to a wealth of solutions, not all of them being optimal [13].

These challenges need to be addressed to further improve strategies for macromolecular complex production, an issue that applies in fact whatever the expression host. These aspects are only starting to be taken into account through the biophysical characterization of the samples at the small-scale level and should help to discriminate the complexes at an early stage of their production based on various biophysical criteria leading to cost-effectiveness and time-saving. It will be essential in the near future to develop specifically these biophysical analyses for macromolecular complexes, once again taking into account the specificities of co-expression approaches, notably with decreased yields and increased heterogeneity.

This chapter addresses these various issues to provide the reader with a broad and comprehensive overview of complex reconstitution by co-expression and characterization in *E. coli*. Specifically, the authors of this chapter have a long-standing expertise in the development and the use of (i) co-expression systems for *E. coli* and (ii) high-throughput protocols dedicated to this host. In this chapter, this expertise has been gathered to provide the reader with an overview of complex reconstitution by co-expression in *E. coli*. Knowledge gained in the last decade on the single protein expression technology using high-throughput methods is provided as starting point and then discussed in the light of the co-expression technology requirements. Importantly, many aspects described here provide specific know-how that is seldom included in publications, explaining that some descriptions are not related to publications but are drawn from our long-term expertise. The complex reconstitution pipeline described in this chapter is provided schematically in Fig. 4.1.

4.2 Getting Started: Choosing the Right Co-expression Strategy

Co-expression technologies can be used in a top-down or a bottom-up manner, depending on the finality and the complexity of the project carried out. In the top-down case, all subunits of the complex are initially co-expressed together in the host to purify the holocomplex. In the bottom-up case, only sub-complexes are initially considered for co-expression experiments and, once these initial sub-complexes are reconstituted, more subunits are added to the co-expression experiments to obtain larger complexes. Yet, whatever the approach followed (top-down or bottom-up), the reconstitution strategy almost always requires an iterative process where the results of the initial experiments are reevaluated to yield new experiments [2, 3, 13, 17].

The choice of using one approach or the other will be dependent on many parameters. For instance, if the complex studied is big, it might be useful to initially consider reconstitution of sub-complexes. Notably, this approach can provide invaluable information on direct protein-protein interactions and complex assembly. Similarly, if the full complex shows intrinsic flexibility, using a more stable sub-complex might be more advantageous. On the other hand, if sub-complexes show poor stability leading to sample heterogeneity, a top-down approach might be more effective. Specifically, all prior knowledge on the complex studied will be of paramount importance in choosing the a priori best strategy, which will then be reevaluated during complex reconstitution.

One often essential aspect of these approaches, like for single protein production, is the use of truncated constructs for the subunits studied. These truncations can help remove regions that might be unstructured and cause poor biochemical behavior of the complexes reconstituted. Such truncations can also be used to separate a complex into more stable sub-complexes, thus also providing knowledge on interacting regions within subunits. Interestingly, removal of some regions may have no effect on the overall stability/formation of a complex, but rather on its function, here again helping decipher between structural and functional modules within the complex subunits. Therefore, protein truncation, even if it may appear as a “loss” when trying to characterize a complex, can provide a wealth of information on the biochemical, structural and functional behavior of this complex.

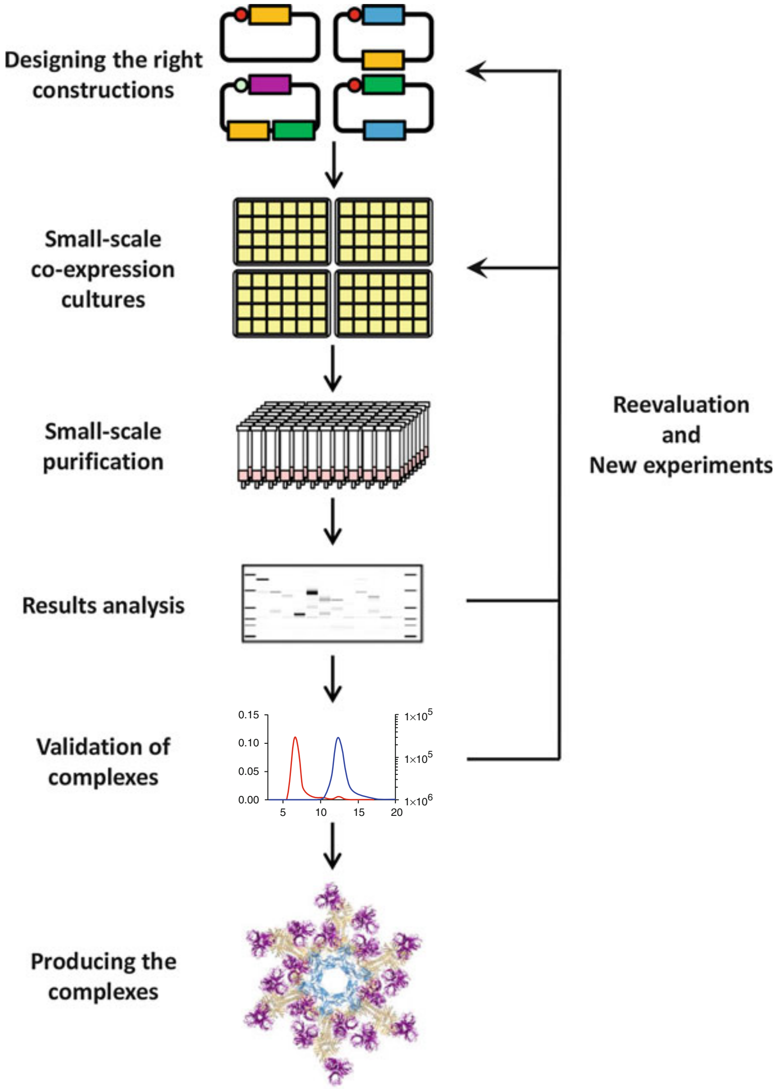


Fig. 4.1 Pipeline of complex reconstitution by co-expression in *Escherichia coli*. The flowchart depicts schematically the series of steps that are carried out in a medium- to high-throughput fashion to reconstitute, produce and characterize complexes through small-scale co-expression experiments in *Escherichia coli*. Complex reconstitution is often an iterative process that requires several cycles of co-expression experiments, results analysis and reevaluation, and design of new co-expression experiments. Once this process has converged and the conditions for complex production have been deciphered, large-scale production can be carried out towards biochemical, structural and functional characterization of the complex. Expression vectors are shown as rounded squares with affinity tags displayed as small circles and the genes of interest as squares. *Colors* indicate different tags and genes. Small-scale cultures in 24-well deep-well plates are shown in *yellow* and affinity resin for retaining complexes is shown in *pink* at the bottom of the 96-column purification deep-well plate. Analysis of the samples by SDS-PAGE and SEC-MALS are depicted. The final structure represents the lactococcal phage TP901-1 baseplate [38]. The same schematic representations are used throughout the figures

4.3 Designing the Right Constructions

Designing the right boundaries for protein constructs is not always straightforward. To help with this, an initial analysis can be performed using a meta server, such as for instance protein CCD [24], that condenses several tools in one analysis. Such a web-based interface collects information (e.g. secondary structure, disordered regions, transmembrane segments, domain linkers) on the protein sequence of interest and helps decide which constructs to be studied. In addition, in the case of single protein expression it has been shown that the use of slightly different boundaries (varying by 5–10 residues) can have important effects on solubility [25]. It is expected that the same effect could be observed with co-expression experiments.

In the case of complexes, although structural modules can be predicted bioinformatically, regions of low complexity may also be extremely important in linking two or more structural domains that are essential for complex stability. Thus, to avoid discarding essential linker regions constructs of various lengths should be considered. These meta-analyses should always be complemented with analyses of sequence alignments done with homologous sequences from organisms covering as much as possible the evolutionary tree of the proteins studied.

Evaluation of co-expression strategies has highlighted major parameters that have to be taken into account when reconstituting complexes [2, 3, 12, 13]. These include the use of one or several plasmids harboring the various genes to be co-expressed, the use of a single or several promoters on each plasmid, and the number of genes under the control of each promoter. At the protein level, the protein(s) bearing the purification tag(s) and the position of the tag at the N- or C-termini of the tagged proteins are absolutely crucial. All these parameters are complex-dependent and have to be evaluated for each project.

Expression without any fusion tag or with a solubilizing protein that cannot bind to an affinity resin can be envisaged, but the use of affinity tags eases considerably the subsequent purification steps and should be preferred. As for single proteins, small affinity tags are preferred such as His-tags or the Strep-tag II [26]. Notably, the small size of these tags is useful as it creates less steric hindrance for complex assembly.

These different findings imply that co-expression strategies require a large number of expression plasmids to be produced. Actually, the issue of the number of clones to be produced is more important in the case of co-expression studies than for single expression. Specifically, the parameters described above imply that every construct will generally be inserted into several expression vectors. In addition, due to the iterative nature of the complex reconstitution strategy that requires that new constructs be made at each iterative step, the number of constructs, and therefore of expression plasmids to be made, is continuously increasing during the whole reconstitution process. Yet, this approach is essential in providing enough flexibility to carry out the co-expression experiments and to progress as rapidly as possible in the reconstitution process.

Collectively, these facts demonstrate that the use of automatable, or at least fast and flexible cloning strategies is of paramount importance to use the co-expression technology to its full potential. The last decade has seen strong developments in cloning strategies that are easily amenable to automation. These methods enable the easy transfer of the genes of interest into multiple vectors regardless of the target sequence and are therefore preferable to a classical restriction/ligation strategy. The alternatives to the "classical" approach rely on recombination, such as Gateway® (Invitrogen) [27] and In-Fusion™ (Clontech) [28], or on ligation-independent cloning strategies, such as LIC and SLIC [29–31]. More recently, RF cloning [32], based on the use of mega-primers and PCR to amplify whole plasmids, expanded even further the possibility to clone multiple blocks in one vector.

4.4 Small-Scale Co-expression Cultures

By generalizing automation and high throughput protocols, the development of structural genomics programs has had a profound effect on the strategy used for performing protein expression in *E. coli*. Notably, these developments lead to the parallelization of protein expression tests through the use of multi-channel pipettes and pipetting robots as well as the use of small-scale cultures in deep-well plates [4, 6–9]. While manual procedures give access to few tens of experiment per weeks, robotics push this to hundreds. These programs have also investigated the importance of various parameters for optimal single protein expression [23, 25, 33–35] that in turn have enabled the identification of initial default expression conditions [36].

One major technological development that has accompanied these changes has been the use of the *E. coli* auto-induction media [37] that enables to get high-density cultures where protein production is automatically started upon shifting from glucose to lactose. In addition, using side-by-side deep-well cultures also minimizes the variables involved in expression, diminishing any artifactual batch-to-batch variations and allowing for a more effective and simplified comparison of results. In most labs, the overall throughput of the expression screening procedures will not be limited by the number of cultures to handle but by the detection system available.

An important parameter to be considered in small-scale expression is the oxygenation of the cultures. Typically, two kinds of deep-well plates can be used for small-scale cultures: 24-well and 96-well deep-well plates. The 24-well deep-well cultures can be grown in most standard shakers without loss of expression. In contrast, with 96-well deep-well plates, the inappropriate oxygenation in these shakers can drastically decrease the protein expression yields and it is therefore required to use specific high-speed and small-orbital shakers [8].

We have previously described protocols for small-scale expression experiments that are almost fully automated and are simple and quick of use [8, 9]. These protocols have been successfully used for the expression of thousands of single proteins, and even for the co-expression of several protein complexes, including 1.8 MDa phage baseplates [38]. Yet, the use of co-expression requires some adaptation from the common protocols used for single protein expression [39].

An essential aspect concerns the inherent combinatorial approach required by co-expression strategies for complex reconstitution. As discussed above, this combinatorial is higher than for single proteins due to the different vectors in which the constructs have to be cloned. Furthermore, a high combinatorial is also required at the small-scale expression level due to the many different combinations of proteins/constructs and vectors encoding them that have to be considered for co-expression. Work on single expression automation and parallelization has provided solutions how to decrease the combinatorial by using basic protocols that have been shown to be applicable to many proteins [9] and that should be transferable, at least in part, to co-expression experiments. Yet, the overall combinatorial of co-expression experiments will always remain higher than for single experiments. In this respect, the use of automated and parallelized protocols should strongly be preferred. Although not available in every lab, pipetting robots are easily accessible in many labs, including to external users. Notably, initial automated protocols for co-expression tests have already been implemented, providing the basis for further, more integrated developments [12, 39].

Another specific change concerns the inoculation of cultures. In the case of single protein expression, the freshly transformed cells are generally used directly, after the 37 °C incubation step, to inoculate the auto-induction media, thus speeding up the small-scale tests [8]. For co-expression, the *E. coli* cells are generally transformed with several vectors and the transformants selected by several antibiotics. Inoculation of the cultures with the fresh transformed cells is therefore sub-optimal and we generally prefer to plate the cells on LB-agar supplemented with the required antibiotics and to let colonies grow overnight. The cultures are then inoculated on the next day by scratching 5–10 colonies from the plate for each small-scale culture.

Finally, an important aspect to take into account is the overall decrease of expression yields observed which is due to the fact that the cells express several proteins at the same time. To address this issue, we favor the use of 24-well rather than 96-well deep-well plates. Specifically, culture volumes of 2–4 mL can be used compared to 0.5–1 mL in 96-well deep-well plates. Related to this topic, we have observed that the auto-induction medium [37] that is generally used for single expression small-scale tests appears less suitable for carrying out co-expression small-scale tests. The reason for this behavior is not very well understood, but we have shown that a specific medium combining auto-induction and use of an inducer (*e.g.*, IPTG) gives better results in small-scale co-expression tests [39].

The rest of the small-scale co-expression culture protocol is very similar to the protocol used for single protein expression [8, 9, 39]. Specifically, these protocols rely on a series of basic conditions optimized for initial protein expression tests, thus decreasing the combinatorial of these initial tests. These various conditions can then be reevaluated in the next experimental rounds (see reevaluation Sect. 4.7 below).

4.5 Small-Scale Purifications

Once cultures are finished, the deep-well plates are centrifuged and the supernatants are discarded. The following stage is then to perform the small-scale purifications. Once again, automated and parallelized protocols are perfectly suited for this analysis and should be preferred since they enable the processing of a larger number of samples and provide better reproducibility at a stage that requires many pipetting steps. Notably, most of these steps are carried out in 96-well deep-well plates that are better suited for purification. Our laboratories have setup conceptually similar protocols for this purification stage [8, 9, 39]. These differ however in some procedures. The differences are due to some specific requirements of the single and co-expression analyses, but not only.

In a first semi-automated protocol [39], most of the steps are carried out through a combination of liquid pipetting (removal, dispense) and centrifugation procedures. At the end of the protocol, no elution from the affinity beads is carried out. Rather, Laemmli buffer is added directly onto the beads prior to analysis. This protocol has the advantage of revealing all the proteins bound to the beads. Indeed, in some cases, during elution from the affinity resin, some proteins or complexes precipitate onto the beads and are not eluted. Such a behavior, which can simply be due to sub-optimal construct size, expression conditions or purification buffer, can prevent the detection of conditions where a complex can be formed. Missing an important, albeit non-optimized, condition could be deleterious for the rest of the project. In a second protocol that is fully automated [8], the purification steps are done in an automated manner using vacuum pumping and a final elution step from the affinity resin. Actually, these two protocols are rather complementary and can be done in parallel or sequentially as they both provided specific information on the proteins/complexes studies.

In both protocols, the first step consists in resuspending the cell pellets in a lysis buffer that can be identical or different to the buffers that will be used subsequently for the purification step. The choice of the lysis/purification buffers is of paramount importance. Specifically, testing various purification buffers has been shown to be very important for single protein production [40] and for complexes [13]. Yet, the use of many different buffers in these initial tests would increase dramatically the combinatorial discussed above, notably when studying complexes. It is therefore better suited in initial co-expression experiments to use a single or a very small set of buffers. Typically, investigating a small set of buffers combining different pHs (e.g., pH 6.0, 7.0, 8.0) and a few salt concentrations (e.g., NaCl 50, 200, 400 mM) should initially be sufficient. Of note, we generally do not add protease inhibitors to check for the stability of the proteins in presence of various partners.

Our protocols rely on a lysis based either on sonication or on enzymatic lysis in presence of large amounts of lysozyme using a few cycles of freezing/thawing the bacteria. There is no reason for preferring one method over the other as they both offer different advantages and disadvantages. Yet, both are interesting since they can also be used during scale-up, keeping the initial conditions used during the small-scale tests. Typically, a volume of 1–2 mL of lysis buffer per test is used to resuspend the cell pellets to avoid working with viscous solutions.

Both protocols make use of batch analysis for the affinity resin despite the possibility of purchasing commercially available purification plates. This helps reducing the price of the analyses, but is also a way to adapt the volume of affinity resin (typically 10–50 μL of beads) to facilitate the concentration of the proteins and their subsequent detection. After transfer of the supernatant/cell lysates onto the beads, incubation is performed during 10–60 min. Note that a too short incubation time can lead to low yield while a too long incubation time can lead to aggregation/precipitation on the resin for some complexes due to the high local complex concentration reached.

4.6 Results Analysis

The samples obtained at the end of the small-scale purification procedure are collected for analysis. Actually, the throughput of the whole expression screening protocol is limited only by the detection system used. Using homemade SDS-PAGE will limit the throughput to a few hundred points/week to reach a maximum of 384 points/week (using a Biorad Dodeca Cell chamber). Using a Labchip GX II system (Perkin Elmer) enables to reach more than a thousand screening point/week with the added advantage of the quantification of the complex subunits yields. For both our purification protocols, all samples collected for gel analysis (*e.g.*, crude extracts, lysis supernatants, washes, affinity resin-bound proteins, elutions) could be analyzed. However, to initially decrease the number of analyses, we typically first load the elution fractions/beads-bound proteins. When some of the results do not meet the expectations (*e.g.*, no complex, some subunits missing), analysis of the corresponding other fractions is carried out to try to possibly decipher the reason of the failure (*e.g.*, proteins not expressed or insoluble).

A successful reconstitution is characterized, after affinity purification, by the retention of all co-expressed proteins on the affinity resin or in the elution fraction, depending on the protocol used. This is a strong indication that the tagged protein has been retained on the affinity resin and that it has also retained the other subunits forming the complex through complex assembly. Of course, it is common that the results observed are not fully matching the expectations. Nevertheless, these results, even if varying from the expected results, provide invaluable information on subunit interactions and complex assembly. Specifically, the use of various combinations of proteins and constructs enables the detailed deciphering of these interactions, strengthening the importance of parallelized co-expression tests. Yet, this analysis of the co-expression tests may not be straightforward in some cases and care should be taken not to over-interpret the data. Several cases can be observed.

A possible complication arises when a low amount is observed for one or several subunits. This might be due to the poor solubility of the protein/construct used. Alternatively, this can arise from the fact that this protein, even if highly produced, is poorly integrated into the complex, either because it misses an important partner or because the buffer conditions are not optimal for an interaction to occur. Deciding between these different possibilities (and possibly others) is not easy. Yet, the use of various constructs/purification buffers can help decipher the cause for this behavior and highlight the importance of thoroughly comparing the co-expression results for retrieving valuable information from them. Importantly, in initial co-expression experiments where the protein constructs and purification buffers have not been optimized, this kind of observation is common and can be essential for the progress of the project. Therefore, even if only a faint band is observed for one or several proteins, one should consider this result carefully and define reevaluation strategies to test its validity: a correct observation may provide the initial track to lead to a successful project.

Another common drawback is the nonspecific binding of some proteins studied to the affinity resin. In some cases, this binding can be quite strong and gives the impression of the formation of a strong complex. Comparison of various combinations of proteins during the co-expression experiments can alert to this problem: a protein always present whatever the co-expressed partners might be nonspecifically bound. Although addition of binding competing agent (*e.g.*, imidazole for his-tag binding affinity resins) to the lysis buffer might be useful to get rid of this interaction, it should be preferred to express the untagged protein on its own and to check its nonspecific binding behavior in absence of its partners.

Another common observation concerns the selective enrichment of some of the proteins compared to the others. Although this might be due to different stoichiometries within the complex, this observation is often due to the enrichment of the tagged protein binding on its own or within a sub-complex. This might be particularly strong when considering many proteins, an issue that is independent of the expression host. This might not always be a problem for co-expression tests, but will complicate subsequent large-scale purification. The use of several proteins bearing different affinity tags with multiple chromatographic steps should improve the analysis, even if the quantity of complex obtained at the end of the purification procedure may be reduced. An important issue will however be the choice of the proteins that should bear the tags.

Although not absolutely important at this stage, the use of multiple constructs for the proteins studied may lead the user to decide to prefer one construct rather than others, depending on the yields observed, in subsequent co-expression tests/large-scale purification. If this is reasonable, one should keep in mind that not all domains of a protein might be important for complex assembly and stability, but can be of paramount importance for function. Therefore, although it might seem interesting to use smaller constructs in subsequent experiments because, for instance, the yields of the complex are higher, all constructs that enable assembly of a soluble complex should be at least listed and whenever possible should be studied in parallel in subsequent characterization/scale-up experiments to select not necessarily the most abundant complex but the one that is the most stable and the closest from the native and functional one (see Sect. 4.8 on the validation of complexes).

4.7 Reevaluation and New Experiments

Reconstitution of a complex through co-expression requires in general, whatever the strategy used, several steps of experimental analysis, results analysis and decisions on new experiments. This iterative reevaluation process, common to single protein projects, is almost mandatory in the case of complex reconstitution by co-expression due to the higher intricacy of this approach. Specifically, the number of parameters that can be modified is large (Fig. 4.2) and the choice of the parameters to be investigated will depend on the results obtained initially.

Change at the protein level is a major parameter to consider. As discussed above, the choice of the right constructs can be of paramount importance for obtaining a complex that behaves correctly biochemically. Finding the correct construct boundaries is not easy and this process might be rather empirical and counterintuitive even if the suggestions given for the design of initial constructs should help ease this work. A good construct may not necessarily be as short as possible, but can also include regions that are poorly conserved and/or apparently not folded. Once again, designing different constructs for the same protein, even on an empirical basis, might prove extremely useful.

The use of fusion proteins to solubilize one or several proteins is another method of choice. This procedure is often investigated during initial

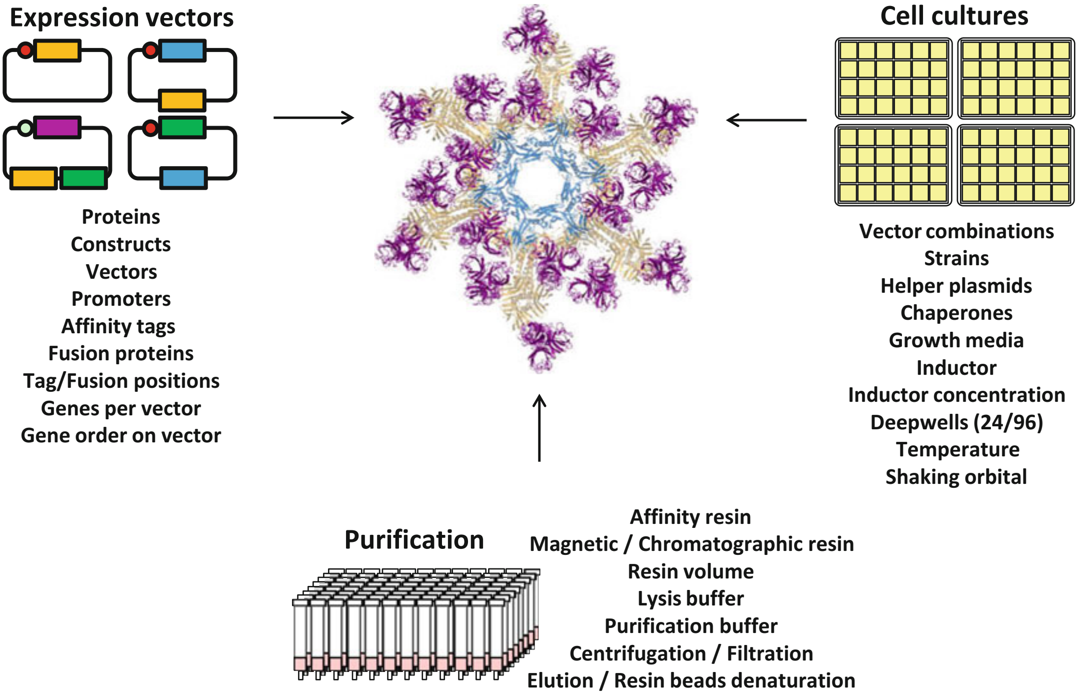


Fig. 4.2 Parameters influencing complex reconstitution and quality. Complex production, yields and quality are influenced by numerous parameters at the vector design, expression, and purification steps. Main parameters are listed on the figure. Some of these parameters should be investigated immediately during initial co-expression experiments. Other parameters should rather be investigated during the reevaluation process. Some of these parameters are common for single protein expression and complex reconstitution by co-expression; others are specifically influencing co-expression experiments. The influence of these different parameters on expression is discussed in the text and in the cited literature

experiments for single protein expression/purification. Co-expression strategies are supposed to prevent or reduce the use of fusion proteins by providing the protein studied with the best solubilizer: its cognate partner(s). This is often what happens and co-solubilization/co-folding is a landmark of co-expression strategies, providing invaluable information on protein/protein interactions [13, 14]. Therefore, use of fusion proteins is generally implemented in reevaluation strategies. Single protein expression evaluations have highlighted particularly useful fusions: thioredoxin (TRX) [41], small ubiquitin-like modifier (SUMO) [42], maltose binding protein (MBP) [43], and glutathione S-transferase (GST) [44]. The interest of these different fusions, and possibly others, in co-expression experiments remains to be investigated.

At the expression level, many parameters can be investigated. Here also, single protein analyses have revealed the importance of several parameters and the requirement of their reevaluation. These include *E. coli* specific strains, helper plasmids, expression media, induction strategies and temperatures [36]. It is expected that these parameters are also important for reevaluation of co-expression experiments and should be further investigated, notwithstanding the fact that other parameters, possibly less important for single protein expression, might be important for co-expression.

Of note, some of the specific strains mentioned contain helper plasmids, such as the ones overexpressing *E. coli* rare tRNAs, to help expressing genes with different codon usage, or those coding for specific chaperones (*e.g.*, GroEL, DsbC). Other helper plasmids however can be used. Specifically, an important aspect for complex reconstitution by co-expression that has been poorly investigated so far is the co-expression of enzymes that will introduce post-translational modifications to favor or even allow complex formation. Such an approach requires that the enzyme depositing the modification is known, but might be of paramount importance in some cases. In addition and as alternative, systems introducing modified amino acids in proteins using specific stop codons are more and more available and provide an ever growing number of modified residues to be used [45–47].

Reevaluation of the lysis/purification conditions should also be carried out. Beside the choice of the lysis method already discussed above, search for optimal buffers is a major parameter to consider. Although this search can be done on small cultures, it may be easier to split a medium culture in small batches to perform all the tests with 96 different lysis and purification buffers in parallel on the same culture (Fig. 4.3). All the lysis/purification steps can be done in 24-well deep-well plates with multi-pipets or can be fully automated. The conditions to screen (*e.g.*, pH, buffer, salts, salt concentrations) can be designed around the initial buffers used through a single double axis experiment or could even take advantage of buffer sparse matrices designed to cover a large multi-dimensional space of purification conditions in a single 96-well experiment. This latter approach would be particularly useful if very no or sub-optimal initial positive buffer conditions are found.

Finally, in a worst-case scenario, initial co-expression experiments might yield only negative results. Considering the numbers of parameters that influence the positive outcome of co-expression experiments, this negative result should not be considered as a reason to stop a project, but rather to carefully reconsider the initial parameters that should be changed to obtain positive results. Considering new proteins, new constructs, other expression and purification conditions, among others, could help deciphering conditions to progress on the project, before deciding to move to different expression hosts that may not prove better solutions, notably if the expression host by itself is not the reason of the failure.

4.8 Validation of Complexes

In current strategies for single protein expression and complex co-expression, small-scale (co-) expression tests are followed by scale-up studies. If the sample can be purified to homogeneity, it is then analyzed biophysically to check its behavior in solution prior to structural and functional analyses. Specifically, a large development of various biophysical methods has been observed in the past decade that enables in-depth biophysical analyses. If these analyses reveal a poor biochemical behavior, other solutions from the small-scale tests are investigated similarly in a reevaluation approach.

The major drawback of this strategy is that even when numerous solutions are provided by the small-scale tests, they will generally not be all analyzed since scale-up studies are less amenable to high-throughput analyses unless specific apparatus (*e.g.*, multi-culture bioreactor) is available. Therefore, empirical decision on the small-scale solution(s) to be further investigated will have to be made based, in general, on the yields of the complex, as determined by SDS-PAGE, without any real validation of the correct biochemical behavior of the native complex. After scale-up and biophysical characterization, if the complex quality is poor, another small-scale solution will be chosen and scaled-up. Such a strategy can be quite time-consuming and costly.

To address this major issue, whenever possible small-scale samples should be directly analyzed biophysically prior to scale-up to help choose the right solutions upfront of scale-up validation. These methods need microgram level of proteins and are quick enough to allow multiple samples to be processed in few hours. A main method is the analytical size exclusion chromatography (SEC) that will characterize the complex presence and its oligomeric state [48, 49]. Whenever possible the coupling of SEC with multi-angle static light scattering (MALS) should be favored [50] to discriminate even further the

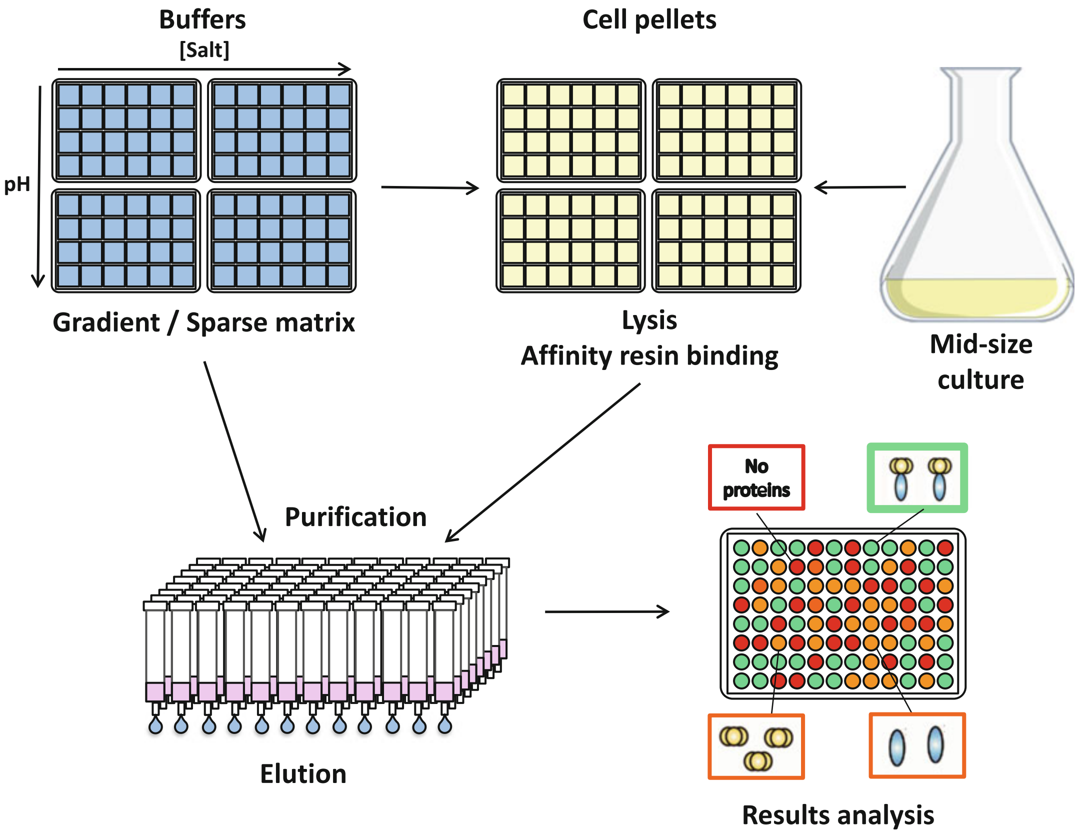


Fig. 4.3 Reevaluation of purification buffers for complex reconstitution. To optimize purification conditions for complex reconstitution, cell pellets from a single mid-size (*e.g.*, 500 mL) culture are lysed and purified under 96 distinct buffer conditions. A single culture is used to avoid small-scale culture batch variation, and pellets from 4 mL of culture are used for each purification buffer. The buffer screens can be designed either to test a few parameters (*e.g.*, pH and salt concentration) as gradients, or to sample a large space of purification parameters through the use of a buffer sparse matrix. In the latter case, it might be useful to make dilutions from sparse matrix stock solutions. The automated protocols described within the text are perfectly adapted for this buffer analysis. At the result analysis step, each buffer will be scored whether it prevents complex purification (*red dots*) or enable partial (*orange dots*) or full (*green dots*) complex formation. These latter conditions will be used to produce the complex at either small or large scale for biophysical validation and further biochemical, structural and functional characterization

mix of sub-complexes and full complexes [51] in the samples. Today, the MALS detectors sensitivity is still rather low compared to the UV detectors, but new generation of machines should fill the gap in a very near future. In addition, differential scanning fluorimetry (DSF) [52] can be used to monitor the stability of the proteins and to discriminate properly folded proteins from soluble aggregates. These assays help to prioritize conditions for scale-up without requiring mid-scale production.

Interestingly, some techniques like native mass spectrometry and electron microscopy have the potential to work with small quantities of samples, even for structural analyses. In addition, it will be interesting to see the developments made around the XFEL (X-ray Free Electron Laser) technology since, potentially, only small quantities of microcrystals might be needed at some point with this technique, possibly only requiring small quantities of samples to be produced.

If biophysical characterization is very important to assess the behavior in solution of a sample, another way to rapidly assess the quality of a protein complex consists in probing its function, whenever a suitable test is available In this case, the quantity of sample required could be small and the functional analysis easily carried out on the samples produced in the small-scale tests. This is of paramount importance since functional modules of proteins may not absolutely be required for complex assembly. Notably, when using different constructs of a protein, some of these constructs may not bear an important functional module. Thus, by coupling biophysical and functional analyses at the small-scale analysis stage, an in-depth characterization of the samples produced can be carried out, highlighting the samples to be used for large-scale production and structural characterization.

4.9 Producing the Complexes

Once the conditions to produce complexes that are amenable to biochemical, biophysical and structural characterization have been identified at the analytical scale, large-scale production should be started to enable further characterization of these complexes in the milligram range. The protocols for small-scale complex reconstitution described in this chapter have been designed to mimic as closely as possible the scale-up conditions. Nevertheless, the move from small-scale to large-scale experiments will in general require some fine-tuning to produce stable complexes.

The required shift between small-scale and large-scale production is generally complex-dependent and requires also in this case empirical analyses through a reevaluation strategy. There are various reasons for the differences observed. One parameter that is important to consider is the culture medium. Although high-density media appear suitable for small-scale tests, this is apparently not always the case during scale-up [39]. The reason for this behavior is unknown and should be further investigated, but probably demonstrates that the small-scale and scale-up culture conditions are not yet exactly the same and would require further optimization (*e.g.*, volume of culture, shaking speed, shaking orbital). Therefore, we tend to use LB or 2xLB media with a specific inducer (*e.g.*, IPTG) rather than auto-induction for complex large-scale expression. Actually, the fact that small-scale expression tests already give better results when using combined auto-induction media and inducer (see Sect. 4.5) appears in agreement with the importance of inducers for complexes.

Another parameter leading to discrepancies between small-scale and large-scale analyses concerns the purification conditions. Specifically, a difference in complex yield and subunit stoichiometry is often the consequence of the variation of the ratio between the volume of lysate and the volume of resin, the overall purification time and the flow rates used during purification. These parameters all have an impact on the local complex concentration on the beads and during the elution. Discrepancies are usually identified at the final gel filtration step. At that stage, to improve complex yield, the buffer for the lysis and purification can be optimized again using micro-dialysis coupled with analytical SEC-MALS [51] and DSF [53]. The same techniques are used to monitor the complex stability on the long term.

4.10 Future Directions

The developments of co-expression technologies in the past decade, whatever the expression host, have strongly facilitated the deciphering of protein complex assembly and the reconstitution of protein complexes for biochemical, biophysical, and structural characterization. In the previous sections, we have described the recent developments of these technologies using *E. coli* as expression host. In addition, we have shown that the developments already made for automated and parallelized deciphering of single protein expression can very often be transposed to co-expression technologies. Yet, the more intricate nature of these latter technologies also requires specific developments.

Specifically, the “next frontier” in developing co-expression technologies appears to be the integration of small-scale biophysical analyses to help choose the right complexes to be used for large-scale production and subsequent biochemical and structural analyses. Once again, developments for single protein production should participate for alleviating this issue. Looking forward, other “frontiers” lay ahead that should also be tackled. We describe some of them in the next last paragraphs as “food for thought”.

One long-standing and broad requirement concerns the assembly of protein/nucleic acids complexes by co-expression. Notably, reconstitution of protein/RNA complexes appears possible and a few cases have been reported [11]. Yet, for many complexes, the binding affinity of the proteins for RNA is often low, which hampers purification of the complexes. It should be investigated whether larger RNA molecules, or protein complexes rather than single subunits, should be used to help form more stable complexes that could be amenable to purification. The feasibility of forming homogeneous extremities for the RNA molecules, possibly by inserting ribozymes on both ends, should also be investigated. Interestingly, small-scale biophysical analyses should help identify correct conditions for forming stable complexes.

With the study of ever larger complexes, the technique of co-expression as it exists will reach some limits. For instance, with an increasing number of subunits co-expressed, yields tend to be lower and, especially, stoichiometry problems become more acute. Actually, this issue is not only restricted to *E. coli*, but concerns to different extents most if not all expression hosts used nowadays. It is therefore important to consider the possibility of performing sub-complexes reconstitution by co-expression techniques to finally assemble the full complex in vitro (see also the Chap. 19 of J. Basquin in this issue). Such an approach might be much more powerful than trying to co-express all subunits at the same time.

Even if *E. coli* might turn out to be less powerful than eukaryotic expression hosts when it comes to express large proteins or introduce specific post-translational modifications, our experience shows that many projects can be successfully tackled using this host. Correct reevaluation strategies are absolutely instrumental in this case. Looking towards the future, we have discussed the current development of producing specific *E. coli* strains that are able to introduce modified amino acid at specific positions in proteins. The development of synthetic biology should further enable the design of even more complicated features and should pave the way for the production of bacteria that should tackle the different issues raised here. Notably, the interest in developing such strains will not only concern researchers working in academia, but will also have essential implications for companies requiring the assembly of macromolecular complexes for specific commercial needs.

Acknowledgments The authors are supported by institutional funds from the Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé et de la Recherche Médicale (INSERM), the Université de Strasbourg, the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01, and by Instruct part of the European Strategy Forum on Research Infrastructures (ESFRI) and through national member agreements.

References

1. Kerrigan JJ, Xie Q, Ames RS, Lu Q (2011) Production of protein complexes via co-expression. Protein Expr Purif 75(1):1–14
2. Perrakis A, Romier C (2008) Assembly of protein complexes by coexpression in prokaryotic and eukaryotic hosts: an overview. Methods Mol Biol 426:247–256
3. Romier C (2008) Protein complexes assembly by multi-expression in bacterial and eukaryotic hosts. In: Sussman JL (ed) Structural proteomics. World Scientific Publishing Co., London, pp 233–250
4. Vincentelli R, Romier C (2013) Expression in Escherichia coli: becoming faster and more complex. Curr Opin Struct Biol 23(3):326–334
5. Barford D, Takagi Y, Schultz P, Berger I (2013) Baculovirus expression: tackling the complexity challenge. Curr Opin Struct Biol 23(3):357–364
6. Almo SC, Garforth SJ, Hillerich BS, Love JD, Seidel RD, Burley SK (2013) Protein production from the structural genomics perspective: achievements and future needs. Curr Opin Struct Biol 23(3):335–344
7. Xiao R, Anderson S, Aramini J, Belote R, Buchwald WA, Ciccosanti C, Conover K, Everett JK, Hamilton K, Huang YJ, Janjua H, Jiang M, Kornhaber GJ, Lee DY, Locke JY, Ma L-C, Maglaqui M, Mao L, Mitra S, Patel D, Rossi P, Sahdev S, Sharma S, Shastry R, Swapna GVT, Tong SN, Wang D, Wang H, Zhao L, Montelione GT, Acton TB (2010) The high-throughput

protein sample production platform of the Northeast Structural Genomics Consortium. J Struct Biol 172(1):21–33

8. Saez NJ, Nozach H, Blemont M, Vincentelli R (2014) High throughput quantitative expression screening and purification applied to recombinant disulfide-rich venom proteins produced in E. coli. J Vis Exp. 2014 Jul 30;(89):e51464
9. Saez NJ, Vincentelli R (2014) High-throughput expression screening and purification of recombinant proteins in E. coli. Methods Mol Biol 1091:33–53
10. An Y, Meresse P, Mas PJ, Hart DJ (2011) CoESPRIT: a library-based construct screening method for identification and expression of soluble protein complexes. PLoS ONE 6(2):e16261
11. Bieniossek C, Nie Y, Frey D, Olieric N, Schaffitzel C, Collinson I, Romier C, Berger P, Richmond TJ, Steinmetz MO, Berger I (2009) Automated unrestricted multigene recombineering for multiprotein complex production. Nat Methods 6(6):447–450
12. Vijayachandran LS, Viola C, Garzoni F, Trowitzsch S, Bieniossek C, Chaillet M, Schaffitzel C, Busso D, Romier C, Poterszman A, Richmond TJ, Berger I (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J Struct Biol 175(2):198–208
13. Diebold M-L, Fribourg S, Koch M, Metzger T, Romier C (2011) Deciphering correct strategies for multiprotein complex assembly by co-expression: application to complexes as large as the histone octamer. J Struct Biol 175(2):178–188
14. Fribourg S, Romier C, Werten S, Gangloff YG, Poterszman A, Moras D (2001) Dissecting the interaction network of multiprotein complexes by pairwise coexpression of subunits in E. coli. J Mol Biol 306(2):363–373
15. Held D, Yaeger K, Novy R (2003) New coexpression vectors for expanded compatibilities in E. coli. Innovations 18:4–6
16. Novy R, Yaeger K, Held D, Mierendorf R (2002) Coexpression of multiple target proteins in E. coli. Innovations 15(2–6):2
17. Romier C, Ben Jelloul M, Albeck S, Buchwald G, Busso D, Celie PHN, Christodoulou E, De Marco V, van Gerwen S, Knipscheer P, Lebbink JH, Notenboom V, Poterszman A, Rochel N, Cohen SX, Unger T, Sussman JL, Moras D, Sixma TK, Perrakis A (2006) Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. Acta Crystallogr D Biol Crystallogr 62(Pt 10):1232–1242
18. Scheich C, Kummel D, Soumailakakis D, Heinemann U, Bussow K (2007) Vectors for co-expression of an unrestricted number of proteins. Nucleic Acids Res 35(6):e43
19. Selleck W, Tan S (2008) Recombinant protein complex expression in E. coli. Curr Protoc Protein Sci. Chapter 5:Unit 5.21
20. Tan S (2001) A modular polycistronic expression system for overexpressing protein complexes in Escherichia coli. Protein Expr Purif 21(1):224–234
21. Tan S, Kern RC, Selleck W (2005) The pST44 polycistronic expression system for producing protein complexes in Escherichia coli. Protein Expr Purif 40(2):385–395
22. Tolia NH, Joshua-Tor L (2006) Strategies for protein coexpression in Escherichia coli. Nat Methods 3(1):55–64
23. Correa A, Oppezzo P (2011) Tuning different expression parameters to achieve soluble recombinant proteins in E. coli: advantages of high-throughput screening. Biotechnol J 6(6):715–730
24. Mooij WTM, Mitsiki E, Perrakis A (2009) ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments. Nucleic Acids Res 37(Web Server issue):402–405
25. Graslund S, Sagemark J, Berglund H, Dahlgren L-G, Flores A, Hammarstrom M, Johansson I, Kotenyova T, Nilsson M, Nordlund P, Weigelt J (2008) The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. Protein Expr Purif 58(2):210–221
26. Schmidt TGM, Skerra A (2007) The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. Nat Protoc 2(6):1528–1535
27. Esposito D, Garvey LA, Chakiath CS (2009) Gateway cloning for protein expression. Methods Mol Biol 498:31–54
28. Berrow NS, Alderton D, Sainsbury S, Nettleship J, Assenberg R, Rahman N, Stuart DI, Owens RJ (2007) A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucleic Acids Res 35(6):e45
29. Aslanidis C, de Jong PJ (1990) Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res 18(20):6069–6074
30. Li MZ, Elledge SJ (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nat Methods 4(3):251–256
31. Jeong J-Y, Yim H-S, Ryu J-Y, Lee HS, Lee J-H, Seen D-S, Kang SG (2012) One-step sequence- and ligation-independent cloning as a rapid and versatile cloning method for functional genomics studies. Appl Environ Microbiol 78(15):5440–5443
32. Unger T, Jacobovitch Y, Dantes A, Bernheim R, Peleg Y (2010) Applications of the Restriction Free (RF) cloning procedure for molecular manipulations and protein expression. J Struct Biol 172(1):34–44
33. Berrow NS, Bussow K, Coutard B, Diprose J, Ekberg M, Folkers GE, Levy N, Lieu V, Owens RJ, Peleg Y, Pinaglia C, Quevillon-Cheruel S, Salim L, Scheich C, Vincentelli R, Busso D (2006) Recombinant protein expression and solubility screening in Escherichia coli: a comparative study. Acta Crystallogr D Biol Crystallogr 62(Pt 10):1218–1226

34. Vera A, Gonzalez-Montalban N, Aris A, Villaverde A (2007) The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. Biotechnol Bioeng 96(6):1101–1106
35. Bird LE (2011) High throughput construction and small scale expression screening of multi-tag vectors in Escherichia coli. Methods 55(1):29–37
36. Vincentelli R, Cimino A, Geerlof A, Kubo A, Satou Y, Cambillau C (2011) High-throughput protein expression screening and purification in Escherichia coli. Methods 55(1):65–72
37. Studier FW (2005) Protein production by auto-induction in high density shaking cultures. Protein Expr Purif 41(1):207–234
38. Veesler D, Spinelli S, Mahony J, Lichiere J, Blangy S, Bricogne G, Legrand P, Ortiz-Lombardia M, Campanacci V, van Sinderen D, Cambillau C (2012) Structure of the phage TP901-1 1.8 MDa baseplate suggests an alternative host adhesion mechanism. Proc Natl Acad Sci U S A 109(23):8954–8958
39. Haffke M, Marek M, Pelosse M, Diebold M-L, Schlattner U, Berger I, Romier C (2015) Characterization and production of protein complexes by co-expression in Escherichia coli. Methods Mol Biol 1261:63–89
40. Graslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, Ming J, Dhe-Paganon S, Park H-W, Savchenko A, Yee A, Edwards A, Vincentelli R, Cambillau C, Kim R, Kim S-H, Rao Z, Shi Y, Terwilliger TC, Kim C-Y, Hung L-W, Waldo GS, Peleg Y, Albeck S, Unger T, Dym O, Prilusky J, Sussman JL, Stevens RC, Lesley SA, Wilson IA, Joachimiak A, Collart F, Dementieva I, Donnelly MI, Eschenfeldt WH, Kim Y, Stols L, Wu R, Zhou M, Burley SK, Emtage JS, Sauder JM, Thompson D, Bain K, Luz J, Gheyi T, Zhang F, Atwell S, Almo SC, Bonanno JB, Fiser A, Swaminathan S, Studier FW, Chance MR, Sali A, Acton TB, Xiao R, Zhao L, Ma LC, Hunt JF, Tong L, Cunningham K, Inouye M, Anderson S, Janjua H, Shastry R, Ho CK, Wang D, Wang H, Jiang M, Montelione GT, Stuart DI, Owens RJ, Daenke S, Schutz A, Heinemann U, Yokoyama S, Bussow K, Gunsalus KC (2008) Protein production and purification. Nat Methods 5(2):135–146
41. LaVallie ER, Lu Z, Diblasio-Smith EA, Collins-Racie LA, McCoy JM (2000) Thioredoxin as a fusion partner for production of soluble recombinant proteins in Escherichia coli. Methods Enzymol 326:322–340
42. Marblestone JG, Edavettal SC, Lim Y, Lim P, Zuo X, Butt TR (2006) Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. Protein Sci 15(1):182–189
43. Kapust RB, Waugh DS (1999) Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. Protein Sci 8(8):1668–1674
44. Smith DB (2000) Generating fusions to glutathione S-transferase for protein studies. Methods Enzymol 326:254–270
45. Lajoie MJ, Rovner AJ, Goodman DB, Aerni H-R, Haimovich AD, Kuznetsov G, Mercer JA, Wang HH, Carr PA, Mosberg JA, Rohland N, Schultz PG, Jacobson JM, Rinehart J, Church GM, Isaacs FJ (2013) Genomically recoded organisms expand biological functions. Science 342(6156):357–360
46. Liu CC, Schultz PG (2010) Adding new chemistries to the genetic code. Annu Rev Biochem 79:413–444
47. Park H-S, Hohn MJ, Umehara T, Guo L-T, Osborne EM, Benner J, Noren CJ, Rinehart J, Soll D (2011) Expanding the genetic code of Escherichia coli with phosphoserine. Science 333(6046):1151–1154
48. Low C, Moberg P, Quistgaard EM, Hedren M, Guettou F, Frauenfeld J, Haneskog L, Nordlund P (2013) High-throughput analytical gel filtration screening of integral membrane proteins for structural studies. Biochim Biophys Acta 1830(6):3497–3508
49. Sala E, de Marco A (2010) Screening optimized protein purification protocols by coupling small-scale expression and mini-size exclusion chromatography. Protein Expr Purif 74(2):231–235
50. Sahin E, Roberts CJ (2012) Size-exclusion chromatography with multi-angle light scattering for elucidating protein aggregation mechanisms. Methods Mol Biol 899:403–423
51. Sciara G, Blangy S, Siponen M, Mc Grath S, van Sinderen D, Tegoni M, Cambillau C, Campanacci V (2008) A topological model of the baseplate of lactococcal phage Tuc 2009. J Biol Chem 283(5):2716–2723
52. Senisterra G, Chau I, Vedadi M (2012) Thermal denaturation assays in chemical biology. Assay Drugs Dev Technol 10(2):128–136
53. Boivin S, Kozak S, Meijers R (2013) Optimization of protein purification and characterization using Thermofluor screens. Protein Expr Purif 91(2):192–206

Membrane Protein Production in *E. coli* for Applications in Drug Discovery

5

Harm Jan (Arjan) Snijder and Jonna Hakulinen

Abstract

Producing high quality purified membrane proteins for structure-based drug design and biophysical assays compatible with typical timelines in drug discovery is a significant challenge. *Escherichia coli* has been an expression host of the utmost importance for soluble proteins and has applications for membrane proteins as well. However, membrane protein overexpression in *E. coli* may lead to toxicity and low yields of functional product. Here, we review the challenges encountered with heterologous overproduction of α-helical membrane proteins in *E. coli* and a range of strategies to overcome them. A detailed protocol is also provided for expression and screening of membrane proteins in *E. coli* using a His-specific fluorescent probe and fluorescent size-exclusion chromatography.

Keywords

Heterologous expression of membrane proteins • Fluorescence size exclusion chromatography

5.1 Introduction

Most of the functional units within cells are formed by macromolecular complexes. The intricate architectural and functional nature of these complexes highlights the challenge faced by researchers in studying these functional players. One of the major bottlenecks of these studies is to reconstitute and purify these complexes as homogeneous samples amenable to biochemical, biophysical and structural characterizations. Of particular importance is the requirement to obtain complexes that display low heterogeneity in terms of composition, structure and function. During early stages of the drug discovery process, access to high quality purified proteins is critical for applications in antibody generation, target validation, in biochemical and biophysical assays, DNA-encoded library screening, determination of the mode of action, *in vivo* studies and in structure based drug design. These different applications require microgram to gram scale

H.J.(A.) Snijder (✉) • J. Hakulinen
Discovery Sciences, AstraZeneca R&D,
SE-43183 Mölndal, Sweden
e-mail: arjan.snijder@astrazeneca.com

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_5

quantities of proteins. The majority of purified proteins required in drug discovery are for human targets, sometimes complemented with orthologs from rodent, dog or another animal model species. However, in drug discovery against infectious diseases, bacterial membrane proteins are of considerable interest. To support these drug discovery activities, we have relied on a range of eukaryotic expression hosts including insect cells, yeast and various mammalian expression systems; nevertheless, the workhorse for soluble protein production remains the bacterial *Escherichia coli* expression system. Bacterial expression has numerous advantages: cost effectiveness, ease of use, short timelines, availability of extensive strain and plasmid collections, and scalability. These advantages are offset by lack of eukaryotic post-translational modifications and processing, toxicity issues of certain target proteins, and for membrane proteins differences in membrane composition. Nevertheless, *E. coli* has successfully been used for the heterologous expression of eukaryotic membrane proteins, including those with relevance to drug discovery, *e.g.*, GPCRs [1–3], MAPEG family members [4, 5] and transporters [6, 7].

To produce membrane proteins in *E. coli* two fundamentally different strategies have been applied. The first strategy aims at producing membrane inserted functional protein and subsequent extraction with retention of functionality. In the other method proteins are produced in aggregated, non-functional form in inclusion bodies, and subsequently refolded *in vitro* to produce functionally folded proteins. Refolding eukaryotic proteins has been reported, and Table 5.1 shows that it is common for structure elucidation using NMR, see [8]. Nevertheless, establishing efficient refolding conditions and generating convincing functional data can be daunting and time consuming.

Before the focus shifts to more practical details in the generation of membrane proteins in *E. coli*, it is essential to describe the process of membrane protein biogenesis and highlight differences between *E. coli* and eukaryotic biogenesis. Almost all α-helical membrane proteins rely on a protein-conducting channel, the translocon, for membrane insertion. The *E. coli* SecYEG complex is located in the bacterial inner membrane, while the eukaryotic counterpart Sec61-complex is located at the endoplasmic reticulum (ER). Structural information of these channels has increased our understanding and helped refine hypotheses around the molecular mechanism of membrane biogenesis [9, 10]. As all proteins, membrane proteins are encoded by mRNA and synthesized at the heart of the ribosome, the growing polypeptide chain extends in ~90 Å long channel through the ribosome towards the exit site. The ribosome with the nascent chain can dock to the translocon [9–12]. The ribosome channel supports formation of a secondary structure in the growing polypeptide chain. At the exit site of the ribosome the nascent chain can interact with proteins docked on to the ribosome. For membrane proteins, the signal recognition particle (SRP) is an essential player, which consists of a 4.5 S RNA and Ffh protein subunit in *E. coli*. The SRP recognizes hydrophobic polypeptide segments, such as those in signal sequences and transmembrane helices. The concomitant conformational change of the SRP allows the ribosome-polypeptide-SRP to interact with FtsY, the signal particle receptor on the membrane. This complex then binds to the SecYEG, to allow co-translational insertion of the membrane protein into the membrane [13]. The current model for folding of membrane proteins into the membrane involves individual insertions of transmembrane helices into the lipid bilayer by transition through a lateral gate opening up in the SecYEG complex, followed by tight packing of the helices into the final functional fold. The packing of the helices into their final functional fold may involve sequential interaction with previously inserted helices, as well as interaction with the translocon and lipids. The general membrane biogenesis mechanism is conserved and shows many similarities between prokaryotic and eukaryotic organisms, as exemplified by successful heterologous overexpression of eukaryotic membrane proteins in *E. coli* [1–6].

Despite the general conservation of the biogenesis mechanism (Fig. 5.1), substantial differences exist that may underlie the challenges

Table 5.1 Examples of structures of eukaryotic membrane proteins that have been produced in *E. coli*

Protein	*E. coli* strain	Vector	Construct	Medium	Detergent	Note	References
CXCR1	BL21	pGEX2a	GST-CXCR1-His	M9	DMPC	Refolded	[8]
TSPO	BL21(DE3)	pET15	–	M9	DPC	SDS purified	[80]
Integrin aIIb3	BL21(DE3)	pMal-C2	MBP-His-TEV-	–	–	Precipitation, organic solvent	[81]
GlyRα1	BL21(DE3) pLysS	pET-31b(+)		M9	LPPG	TM domain only	[82]
α2β2 nAChR	Rosetta 2(DE3) pLysS	–	His tagged	M9	LDAO	TM domain only	[69]
α7 nAChR	Rosetta 2(DE3) pLysS	–	His tagged	M9	LDAO	TM domain only	[48]
UCP2	Rosetta(DE3)	pET21	UCP2-His6	–	TX100/DPC		[83]
FXYD1	C41(DE3)	pBCL173/99	Bcl-XL-FXYD1	M9	SDS	Purification under denaturing conditions	[84]
phospholamban	BL21(DE3)	pMalc2x	MBP-PLN	M9	DPC	Denaturing purification	[85]
Synaptobrevin	BL21(DE3) or BL21(DE3) RIL	pET28	His6-TCR-Snare	M9	Sodium cholate, DPC		[86]
mPGES-1	BL21(DE3) pLysS	pSP19T7LT	His6-mPGES	TB	Red TX100		[4]
FLAP	BL21(DE3)	pET28a	FLAP-His	–	DDM/C12E8-C8E4		[87]
Cyt b561 B	BL21(DE3)	pET15B	His6-TCR-cyt	–	DM/NG		[88]
ATM1	BL21	pASK-IBA1	SP OMPA-His8-ATM1-StrepII	LB	DDM		[7]
rNTS1	BL21 Tuner	pBR322-der	MBP-His-3C-NTR1-3C--TrxA	2TY	DM/CHAPS/CHS, DM, NG		[1]
PfAQP	CD43(DE3)	pET28	His6-TCR-AQP	LB	BOG		[89]

(continued)

Table 5.1 (continued)

Protein	*E. coli* strain	Vector	Construct	Medium	Detergent	Note	References
Syntaxin 1A7SNAP-25/ synaptobrevin-2 Neuronal SNARE complex	BL21(DE3)	pET28	His6-TCR-Snare	–	NG		[68]
Kir3.1(GIRK1)	BL21(DE3)	pET28	His6-TCR-pore-TCR-His6	LB	DDM/NG	Chimeric	[90]

BOG n-octyl- β-D-glucopyranoside, *DPC* dodecylphosphocholine, *DM* n-Decyl β-D-maltopyranoside, *DDM* n-Dodecyl β-D-maltopyranoside, *LPPG* lyso-palmitoyl phosphoglycerol, *LDAO* Lauryldimethylamine-oxide, *NG* n-nonyl- β-D-glucopyranoside, *SDS* Sodium dodecylamine oxide, *TX100* Triton X100

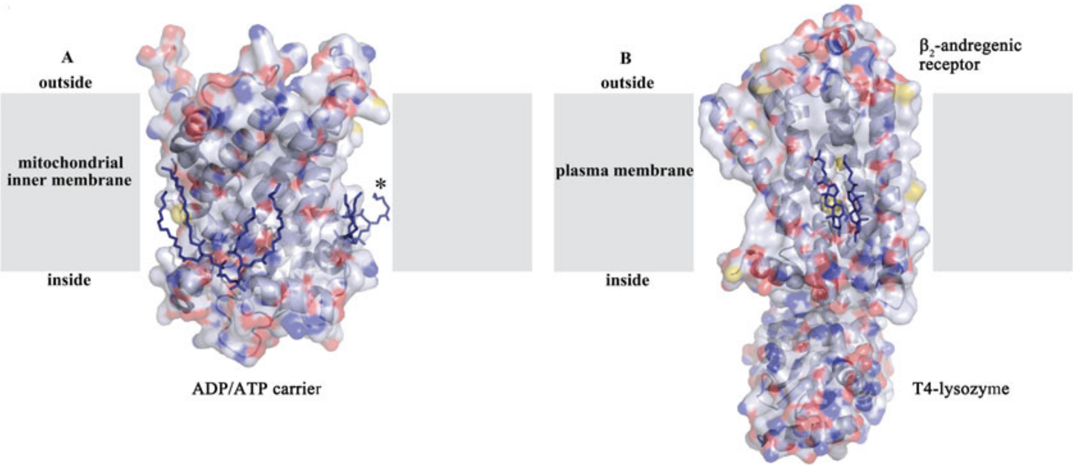


Fig. 5.1 Lipid binding in two membrane protein structures. (**a**) A surface/cartoon presentation of mitochondrial ADP/ATP carrier structure [37] with three bound cardiolipin molecules. Cardiolipin marked with an asterisk is involved in stabilization of interactions between two monomers. (**b**) A surface/cartoon presentation of the β2-andrenergic receptor structure with T4-lysozyme [39]. Structurally important cholesterol molecules are bound between helices I, II, III and IV. Lipid molecules are shown in deep blue (The figure was created using PyMOL (DeLanoScientific LLC))

observed during heterologous expression of eukaryotic membrane proteins in prokaryotic hosts. A full description of all differences goes beyond the scope of this chapter, but a number of key differences are highlighted, which all have implications for membrane protein biogenesis. First of all the translation/elongation rates differ between prokaryotic and eukaryotic organisms, with in *E. coli* polypeptide elongation rates vary from 10 to 20 amino acids per second depending on growth rate and growth conditions [14]. In eukaryotes elongation rates are considerably slower, in the range of three to eight amino acids per second (references thereof in [15]). These differences in elongation rate may affect subsequent folding and membrane insertion.

In addition to the different elongation rates, mechanisms of translational pausing, and arrest and direction to the translocon differ between eukaryotes and Gram-negative bacteria; the latter lack the Alu domain in the SRP that causes translational arrest until engaged with the translocon. Recently, Li, Oh and Weissman [16] showed that translational pausing was achieved in *E. coli* by internal Shine-Dalgarno sequences. An enrichment of such pause sequences was observed before the second transmembrane helix, a translation phase in which membrane proteins are assumed to be targeted to the translocon [17]. Another marked difference between prokaryotes and eukaryotes is the ratio between ribosomes and signal recognition particles, prokaryotes have about an order of magnitude fewer SRP per ribosome compared to eukaryotic cells [18, 19].

After co-translational insertion into the endoplasmic reticulum, eukaryotic membrane proteins undergo a complex sorting and maturation process through the ER and Golgi apparatus. The process involves glycosylation, proteolytic processing and includes mechanisms for quality assurance. Prokaryotes lack similar refinement in membrane protein biogenesis, with limited internal membrane structure and limited membrane composition divergence.

Both eukaryotes and prokaryotes appear to posses limited amounts of membrane protein biogenesis machinery that can become saturated. The current consensus is that this overloading of the biogenesis machinery is toxic to cells in situations of (heterologous) overproduction of membrane proteins [20], although the biological function of the overexpressed protein may contribute to toxicity as well. All the differences highlighted above between eukaryotes and *E. coli* production host can contribute to saturate and misbalance the biogenesis machinery and thus pose particular challenges for heterologous expression of membrane proteins in *E. coli*. Reducing these adverse effects have led to improvement of membrane protein expression. A range of strategies has been applied, some specifically designed to reduce stress on biogenesis machinery [21–23], other more empirical strategies include optimization of promoters, strain development [24], culture and induction conditions [25], construct design, co-expression of chaperones or translocation machinery [26] and combinations of the above. With further appreciation of the intricate details of membrane biogenesis, more improvements to heterologous expression of membrane proteins in *E. coli* are to be expected.

Another challenge for heterologous overexpression and production of eukaryotic membrane proteins in bacterial hosts, resides in differences in the lipid bilayer. Table 5.2 illustrates prominent differences in membrane composition between *E. coli* and eukaryotic organisms; for instance, *E. coli* lacks sterols, sphingolipids and poly-unsaturated fatty acids. In addition, phospholipid head group composition differs substantially between different organisms. In turn, eukaryotes have greater variety of different membranes, each with their own characteristic lipid compositions and associated biophysical properties. As eukaryotic membrane proteins progress through the secretory pathway, they are exposed to different lipid environments, *e.g.*, membrane thickness was shown to vary from 37.5 Å in the ER, 39.5 Å in Golgi membranes, and 35.6 Å and 42.5 Å in the basolateral and apical membranes of rat hepatocytes [27]. The thickness was shown to be modulated by embedded proteins whereas cholesterol had a minor effect. A recent computational analysis showed that the hydrophobic thickness for α-helical membrane proteins from bacteria and eukaryotic ER is ~30–31 Å, for eukaryotic plasma membrane proteins ~30–37 Å, for mitochondrial inner membrane proteins ~27–30 Å and for β-barrel proteins from bacterial and mitochondrial outer membrane proteins ~23–25 Å [28]. Thus, hydrophobic mismatch may occur between protein and membrane. Consequences of hydrophobic mismatch were investigated in a model system with *E. coli* MelB where maximum activity was obtained when the hydrophobic thickness of the membrane and that of the protein were matched [29]. Such a hydrophobic mismatch potentially could be more severe for eukaryotic proteins heterologously expressed in bacteria.

Lipid-protein interactions are highly dynamic with exchange rates of ~2×10^7 s^{-1} [30, 31]. Bulk lipids, *i.e.*, not in direct interaction with proteins, affect membrane proteins through lateral pressure, membrane tension, curvature and fluidity [32]. More specific interactions between lipids and proteins can occur as well, with lipids binding into clefts and cavities in protein or multisubunit complexes (Fig. 5.1). They can be also located within proteins and function there as cofactors [32]. Specific lipids are essential for the function of several eukaryotic membrane proteins either modulating the membrane or directly by interacting with the protein [33]. Mammalian Na^+/H^+ antiporter NHE3 is located in membrane domains with cholesterol and sphingolipid, and modulation of its activity requires electrostatic interaction with anionic lipids [34]. There are strong indications that lipids are important for proper insertion, folding and topology of membrane proteins [35, 36]. Several membrane protein structures have been elucidated with bound lipid molecules. Here a few examples are mentioned. The bovine ADP/ATP carrier crystallized as a dimer with protein-protein interactions mediated by a cardiolipin molecule tightly bound between the two monomers (Fig. 5.1a) [37]. Cardiolipin has been found also on interfaces of protein complexes of the respiratory chain, such as cytochrome bc1 and cytochrome c oxidase complexes; cardiolipin has been proposed to play a pivotal role in assembly of these supercomplexes [38]. Similarly, β2-adrenergic receptor crystallized as a dimer with two cholesterol molecules bound per monomer (Fig. 5.1b). Cholesterol was not at the interface of the two receptors, but it was specifically binding in a surface groove between helices I, II, III and IV. Molecular simulations showed that choles-

Table 5.2 Lipid composition of different cell types as %

	PC	PE	PG	CL	PS	PI	PA	SM	LPC	PS+PI	Others
E. coli[a]	–	70–80	20–25	>10	–	–	–	–	–	–	
S. cerevisiae plasma membrane[b]	16.8	20.3	–	0.2	33.6	17.7	3.9	–	–	–	6.9
S. cerevisiae mitochondria[b]	40.2	26.5	–	13.3	3.0	14.6	2.4	–	–	–	ND
Dendritic cells exosoms[c]	43	23	–	–			–	9	13	12	
BHK21[d] plasma membrane	26	29	–	–	18	3	–	24	–	21	

PC phosphatidylcholine, *PE* phosphatidylethanolamine, *PG* phosphatidylglycerol, *CL* cardiolipin, *PS* phosphatidylserine, *PI* phosphatidylinositol, *PA* phosphatidic acid, *SM* sphingomyelins, *LPC* lysophosphatidylcholines, *ND* not detectable

[a] Dowhan (1997)

[b] Zinser et al. (1991)

[c] Subra et al. (2007)

[d] Opekarova and Tanner (2003)

terol is important for increasing the packing of helices II and IV and stabilizing the receptor [39]. Incompatibility of lipids necessary for membrane biogenesis, folding or function may underlie further challenges to obtain and purify functionally active eukaryotic membrane proteins in bacterial host systems. Nevertheless, the *E. coli* expression systems should be considered in a bioreagent generation strategy and evaluation of its applicability could be well worth the effort.

5.2 Expression Strains and Promoter Systems

The *E. coli* BL21(DE3) strains are widely used as recombinant expression systems. The overexpression in these strains is driven by the T7 RNA polymerase, whose expression is regulated by the *lac*UV5 promoter. The *lac*UV5 promoter is a stronger promoter than the natural *lac* promoter [40]. Table 5.1 lists *E. coli* expressed eukaryotic membrane proteins for which high resolution structural information has been obtained. The table illustrates that BL21(DE3) has been applied successfully in a number of instances to obtain high-quality membrane proteins. The system does suffer from some disadvantages, such as the fact that the Lac repressor mechanism is not tightly regulated and it is a poorly titratable system. Basal expression of T7 RNA polymerase leads to continuous production of the target protein, which may be toxic to the growing *E. coli* cell even prior to induction. Studier used coexpression of T7 lysozyme to inhibit polymerase activity and to repress basal levels of target protein production [41]. This strain, BL21(DE3) pLys, is particularly useful for expression of toxic proteins. Krepkiy compared BL21(DE3), DH5α and AVB101 *E. coli* strains for the expression of the cannabinoid receptor CB2 [42]. Of those, BL21(DE3) resulted in highest production of functional receptor.

Miroux and Walker derived C41 and C43 variants from BL21(DE3) which supported improved expression of ATP synthase subunit b [24]. The over-production was concomitant with formation of intracellular membranes rich in b subunit [43]. These strains are now widely used for production of other membrane proteins. Detailed analysis of these strains showed mutations in the *lac* UV5 promoter resulting in reduced levels of T7 RNA polymerase compared to the parent BL21(DE3) system [21]. Lower levels of T7 polymerase results in fewer copies of mRNA of the target protein, which is hypothesized to reduce the load on the translocation machinery and hence lead to improved production of membrane proteins. De Gier and co-workers then developed a derivative strain of BL21(DE3), the Lemo21 strain, in which T7 lysozyme, the natural inhibitor of T7 polymerase, is expressed from a tightly regulated rhamnose promoter. Thus, T7 polymerase activity can be reduced by titrating in rhamnose to optimize sustained bacterial cell growth during expression of membrane proteins, thereby reducing the stress response and toxicity and maximizing the yield of functional membrane proteins. For different membrane proteins, different concentrations of rhamnose were required for optimal membrane protein production, highlighting the general usefulness of the Lemo21 strain [22]. *E. coli* BL21-AI uses a different solution to match transcription and membrane protein biogenesis capacity, whereby T7 polymerase is controlled by the tightly regulated arabinose promoter, while the target gene is under the control of the T7lac promoter. Both arabinose and IPTG induction ensures full target gene transcription. Under restrained expression conditions, protein expression is induced by arabinose only and relies on sporadic dissociation of LacI from the LacO1 repressor for transcription of the target gene to occur. A series of 30 prokaryotic homologs of cardiac Na^+/Ca^{2+} exchangers were produced in 5–25-fold larger quantities compared to IPTG induced BL21(DE3) [23]. The BL21 Tuner™ cell strains are deficient in *lacZY*, which enable adjustable levels of protein expression throughout all cells in a culture. IPTG enters *E. coli* independently of the permease pathway. This, in turn, allows uniform and concentration dependent regulation of the expression. Again, lower expression levels may support correct membrane insertion, and reduced toxicity. The system has been applied by, *e.g.*, Egloff [1] for expression of Neurotensin Receptor R1. Improved membrane protein overproduction was achieved in strain EXP-Rv1337-4, which showed a decrease in plasmid copy number. This reduction in copy number likely slows expression to a level that the translocation machinery can accommodate [44]. Further alternatives to reduce the translocation machinery load may be achieved using the *E. coli* Sm^P strain, *i.e.*, CH184; this strain harbors mutations in the S12 protein (streptomycin resistance protein) of the ribosome which results in a hyperaccurate but slow translation phenotype (~5 aa/s). By subsequently titrating in streptomycin the phenotype can be relieved. The reduced translational speed has been shown to enhance eukaryotic protein folding efficiency [15]. As far as we are aware, this mechanism has not been exploited to improve eukaryotic membrane protein expression, but may well be an efficient tool to reduce toxicity and translocational load.

Although basal transcription is being exploited in the restrained expression above [23], for specific target genes basal expression still resulted in toxicity [18]. A reduction in basal expression has been achieved by combining phoA and Lac promoters (phac promoter) and further combining that with a lambda t0 transcriptional terminator upstream of the phac promoter (tphac). Moving from phoA to a tphac promoter for expression of two GPCRs, toxicity and basal expression were reduced and the quality of the purified receptors was improved [18].

Membrane protein expression has been improved by an alternative approach based on alleviating the mismatch between transcription and membrane biogenesis. Nannenga and Baneyx [26] showed that inactivation of trigger factor and coexpression with YidC significantly improved production of multi-topic membrane proteins in the inner membrane of *E. coli*. Trigger factor competes with the SRP on the ribosome for nascent polypeptides and its deletion may lead to increased numbers of ribosomes translating SRP-dependent substrates. In another study, the positive effect of YidC on expression of GPCR's in *E. coli* was not confirmed [45], but significant improvement was observed with overexpression of FtsH, a membrane-anchored AAA+ protease. The full rationale for improved GPCR expression with the overexpression of FtsH is not yet understood, but changes in lipid composition were suggested.

5.3 Construct Design

While the choice of expression strains is important, construct design probably requires most consideration to ensure success. Nevertheless, the design process requires a fair bit of trial-and-error and cycles of design and evaluation. Constructs need to be designed for transcription of the genetic code to mRNA, translation into a polypeptide chain, followed by translocation into a functionally inserted membrane protein. Tags need to be designed, which allow detection and purification without affecting functionality of the protein. The actual protein may require modification, such as sequence truncations and site specific mutations, to enhance stability or homogeneity. These choices are largely dictated by downstream applications of the purified membrane protein, where structural biology typically is the most demanding with respect to quantities, stability and homogeneity.

Over the recent years much of the traditional molecular biology work in construct generation has been replaced by commercially available DNA synthesis services. Proprietary algorithms are applied to optimize DNA sequences toward codon usage, GC-content, sequence motifs and RNA secondary structure. However, membrane protein biogenesis has been shown to be codon-sensitive, with single synonymous codon substitutions influencing mRNA stability and structure, but also influencing translational initiation, elongation and protein folding and translocation (reviewed in [46]). Effects of codon optimization on membrane protein expression are contradictive, with both positive and negative effects reported on yields of functionally active overexpressed membrane protein [22]. This reflects either target specific effects, variations in optimization algorithms applied or combinations of these two. Above all, it highlights that further understanding and research is require to fully exploit the RNA code for the optimization of heterologous expression of membrane proteins.

In a recent report, specific synonymous Ser substitutions to the 5′ coding region adjacent to the AUG start codon were sufficient to significantly improve expression of two *E. coli* membrane proteins in *E. coli* [47]. The authors speculate that the changes affected translation initiation, making initiation more efficient and thus improving expression. In a study where human membrane proteins were overexpressed in *E. coli*, the translation initiation was examined in detail using varying leader sequences. The translation initiation rate was shown to be the crucial variable in the expression of CD20 and EG-VEGFR1, with weaker translational strength supporting higher membrane protein expression possibly due to reduced overload of the translocation machinery [18]. Whether translation initiation requires up or down regulation for optimization of membrane protein expression is presumably target and growth condition dependent, and in cases of low or modest membrane protein expression codon choice requires careful consideration.

Modifications to the protein coding sequence are aimed at generating better behaving proteins during purification and in final applications. The final application dictates what is required while the biological function and pharmacology, on the other hand, dictate what modifications are still acceptable. Structural studies require typically pure, concentrated, and homogeneous protein preparations, which are stable over long time spans. Mobile regions may need to be removed for crystallization (for example in [1]) or separate domains may need to be studied to keep total size within manageable ranges for NMR structure determination [48]. Additionally, for structure determination often demands are put on the detergent system, *e.g.*, small and deformable micelles, which typically are not supporting optimal stability for membrane proteins. Individual point mutations and combinations of point mutations can be introduced to increase stability of membrane proteins, either in systematic ways [49, 50] or by evolutionary optimization screens [51–53] and even using computational methods [54]. These methods have been successful in obtaining structural information from membrane proteins, but still require careful pharmacological characterization of the modified proteins.

Introduction of stable folding soluble domains in exposed loops of membrane proteins has become another important tool in membrane protein structural biology. This strategy has been reviewed in [55]. Various folding domains such as T4L exert a stabilizing effect on the membrane protein when inserted at the correct location. The exposed soluble domain may mediate crystal contact formation. This method has been applied for GPCR's mostly using insect cell expression systems, but it would also be applicable for the heterologous expression of membrane proteins in *E. coli*.

In the choice of purification tags, the His tag is ubiquitously used as first capture step after solubilization. The length of the tag as well as location may influence expression and purification properties. Longer deca-histidine tags improve capturing and allow more stringent washing conditions and hence are commonly used for membrane proteins that do not express to high levels [56]. Mohanty and Wiener [57] observed for a tetrameric aquaporin a deca-histidine tag resulted in lower yield and increased aggregation of the purified aquaporin compared to a hexa-his tag. In the same study, N-terminal tags were slightly favorable over C-terminal tags. A similar bias was observed in an overexpression screen of 24 membrane proteins from *Legionella pneumophila* in *E. coli* [25]. A bias for C-terminal histidine tags was observed in a different study on 48 different prokaryotic membrane proteins. For a number of targets expression depended crucially on the location of the tag [58]. The observed preferences for tag location and differences observed in these studies may originate from changes in the 5′ coding region in constructs, the importance of which was highlighted in [47].

An optimized system for expression of GPCRs in *E. coli* has been developed by Grisshammer and co-workers, and it has been applied for mg scale production of Neurotensin receptor, the A2a receptor and CB2 receptor [2, 3, 42]. The GPCRs are expressed with an N-terminal maltose binding protein, which facilitates efficient translocation of the N-terminal part of the GPCR. The receptors were further decorated with thioredoxin A (TrxA) and either His, biotinylation recognition sequences, FLAG tags or combinations of these. TrxA was shown to have a stabilizing effect and increased levels of pure protein.

A number of strategies apply membrane specific fusion proteins; MISTIC is a bacterial membrane associated protein which autonomously folds into the membrane. When fused to the C terminus, diverse classes of eukaryotic membrane proteins could be produced in a membrane-embedded form [59], including six GPCRs, three potassium channels and seven TGF-receptors. The CB2 receptor has been expressed as an N-terminal MISTIC fusion in combination with a C-terminal TarCF fusion, achieving a receptor density of close to 1 pmol/mg membrane protein [60]. MISTIC has also been successful to produce a plant ADP/ATP carrier; here, the MISTIC fusion showed lower activity and tendency to form large oligomeric structures, however full activity and correct oligomeric state was recovered after fusion protein cleavage [61]. The well expressing GlpF, an *E. coli* glycerol channel, has been used as N-terminal fusion partner to improve expression of three human proteins in *E. coli* in membrane inserted from. Expression of these GlpF-fusions in *E. coli* was not associated with the cytotoxicity normally associated with high-level membrane protein expression. Due to the GlpF topology, this strategy is limited to membrane proteins with an intracellular N-terminus [62].

5.4 Growth and Induction

Table 5.1 illustrates that standard medium (*i.e.*, LB and M9 medium) is a common choice for expression of membrane proteins. Richer media such as 2TY and TB have been applied as well. In one study where the composition of the 2TY medium was modified, higher levels of peptone were shown to be detrimental to growth, while reduction of peptone to 0.8 % (w/v) with a concomitant increase of yeast extract to 3.2 % (w/v) was optimal for membrane protein production [25]. The NPS medium [63] was also observed to support high-level membrane protein expression [25]. Whereas these richer media may not necessarily lead to higher expression per cell, increase in biomass results in higher protein yield per liter.

Careful optimization of inducer concentration will be required to obtain high-level protein expression and membrane insertion (see Fig. 5.2), irrespective of induction method and promoter/regulator choice [22]. Should one want to reduce the number of parameters to screen then, in general, lower inducer concentrations are recommended. A control without any inducer is advisable and leakiness of promoter may be sufficient to support expression of functional membrane protein [23]. Autoinduction using α-lactose as inducer has been successful for soluble proteins [63] and also for membrane proteins, where autoinduction has been shown to be an attractive induction method [25]. The superiority of autoinduction may reflect adaptation of bacteria to deal with gradually increasing load on the membrane biogenesis machinery.

Growth temperature is another key parameter and often a reduction from 37 to 18–20 °C has resulted in production of higher levels of functional membrane protein. A multitude of cellular changes occur in response to decreased temperature, including changes in membrane composition, induction of cold shock chaperones, reductions in translational speed and changes in RNA [64]. All these changes may lead to reduced toxic effects from membrane protein expression, but thus far no study has conclusively deconvoluted the individual contribution of those factors to the observed beneficial effects. In analogy with chaperone induction by cold-shock adaptation, several reports have described successful coexpression of membrane proteins with chaperones, *e.g.*, DnaJ/K, GroEL/ES. Chaperones have not had a positive effect in all cases, and advantageous effects of chaperones may be highest under circumstances when protein production, membrane targeting and membrane translocation are not balanced; conversely, in the case of strict co-translational membrane protein biogenesis, the effect of chaperones may be negligible.

5.5 Expression Screening

It is apparent that sampling of a large number of constructs, host strains and conditions is required. This, in turn, requires efficient, rapid, and sensitive expression screening methods with a minimum of manual handling. The goal of an expression screen is to evaluate which constructs, strains, and growth conditions give optimal levels of functional protein. While for soluble proteins immunoblotting is a suitable tool, for membrane proteins immunoblotting is inadequate since a large fraction of the produced membrane protein may be in non-functional states. Indeed, in a comparison of 100 GPCRs heterologously expressed in *E. coli*, yeast and mammalian cells, no evident correlation between immunoblot signals and GPCR ligand binding capacity was observed. Particularly, all 46 receptors that showed immunoblot signals in *E. coli* were produced in non-functional form [65]. Instead, functionally membrane-inserted protein ought to guide decisions on constructs and expression conditions. Probing functionality of membrane

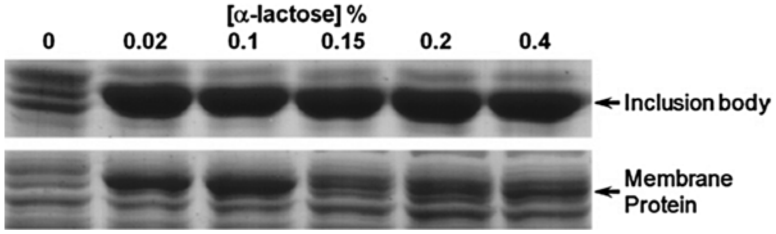


Fig. 5.2 The effect of inducer concentration on the overproduction of two membrane proteins in *E. coli*. The *upper gel* shows an outer membrane protein target produced as inclusion bodies. The *lower gel* shows the expression of a 93 kDa P-type ATPase. Titration of inducer concentration is required to obtain optimal expression in the membrane fraction (The figure is reproduced from Ref. [25] with permission from Elsevier Inc)

proteins has not always been feasible due to lack of tools, *e.g.*, radioligands, or in those cases where function may not be known up-front. Membrane insertion can be tested by membrane preparation protocols; however, multiple membrane preparations at small scale are poorly accurate, time consuming and labor intensive. Biochemical behavior and aggregation state have been used as quality criteria for membrane protein expression and surrogate for membrane insertion [58]. Fluorescence-detected Size-Exclusion Chromatography (FSEC) has emerged as a highly sensitive, fast method for expression screening of membrane proteins [66]. The method involves fusing a fluorescent protein (*e.g.*, GFP) to the target for the detection in crude lysate. The fluorescent signal informs on total extractable protein levels, but more importantly the chromatographic profiles contain information on protein quality and oligomerization state. While FSEC has revolutionized membrane protein screening, the GFP fusion strategy has some disadvantages, including false positives and requirement for proteolytic removal or recloning of constructs. FSEC using a small fluorescent histidine specific probe circumvents many disadvantages associated with GFP based FSEC [67]. Since the probe uses chemistry similar to that applied in IMAC purification, when using this small molecule in FSEC applications additional information is collected that helps designing purification strategies.

5.6 Multisubunit Complexes

While heterologous expression of membrane proteins is not a trivial task, generation of heteromeric multisubunit complexes presents even greater challenges. Nevertheless, in some cases, complexes have been obtained by stoichiometric mixing of purified separate subunits [68, 69]. Bondarenko showed by MALS and NMR formation of a pentameric transmembrane domain of an nAChR channel from separate α4 and β2 subunits. Most functional multimeric complex may not be produced in this fashion, and in fact generation of fully intact α4β2 nAChR channel including the extracellular domain was not possible by mixing the subunits [69]. Instead coexpression may be required, as exemplified by the *Vibrio cholera* TRAP-transporter (tripartite ATP-independent periplasmic). Subunits Q and M had to be coexpressed as dicistronic gene containing the native intergenic region between siaQ and siaM to obtain a functional transporter [70]. Furthermore expression of multisubunit complexes can be restricted by one of the subunits, as in the case of bacterial F-type ATPase consisting of eight different subunits ($ab_2c_{10}\alpha_3\beta_3\gamma\delta\varepsilon$). Based on several mutational studies and time-delayed *in vivo* assembly assays, it has been shown that subunits ab_2 and $c_{10}\alpha_3\beta_3\gamma\varepsilon$ form stable subcomplexes, before they are assembled by subunit δ into a functional ATP synthase (reviewed in [71]). Several studies have shown that overexpression of the transmembrane subunit a (*uncB* gene) causes growth inhibition of *E. coli* [72]. Degradation and instability of the *uncB* mRNA and folding, insertion and proteolytic digestion of subunit a control the expression of F_o [73–75]. In contrast, the two other F_o subunits (c and b_2) can be expressed individually in high yields [76]. Membrane insertion of the a-subunit is dependent on membrane-embedded subunits b and/or c [77, 78].

Co-purification and functionality has been used to confirm complex formation, but the FSEC screening methodology described above can easily be extended for the analysis of multimeric complexes [79].

5.7 Outlook

Escherichia coli is an expression host of paramount importance for soluble proteins and has been used for membrane protein expression as well. The advantages of *E. coli* as host system are rapid replication and short timelines, low-cost, easy scale-up, and a well-established toolbox for genetic manipulations. *E. coli* has been used for heterologous expression of eukaryotic membrane proteins and should be considered along other expression systems such as yeast, insect cell and mammalian cells. With a further increased insight in membrane biogenesis in both pro- and eukaryotes, more successful cases of overproduction in *E. coli* may become available.

5.8 Protocol for Plate-Based FSEC Expression Screening

5.8.1 General Considerations

This protocol describes small-scale expression of membrane proteins in *E. coli* and analysis by FSEC using a His-tag specific fluorescent probe. The protocol is applicable for FSEC analysis of GFP-tagged proteins, in which case the His-tag specific fluorescent probe can be omitted. The protocol is based on 3 mL culture volumes in 24 deep-well plates, but can be adapted to 96 deep-well plates.

Growth can be analyzed in detail by measuring optical densities; use a multipipette and transfer 50 μL of culture medium to a 96 well flat-bottom plate (Corning), ensuring that no air bubbles disturb the meniscus. Measure absorbance in a plate reader, *e.g.*, BMG FLUOstar reader in absorption mode with a 405 nm filter with 10 nm bandwidth. Conventional optical densities (A_{600}) are tenfold higher than the measured values. For different readers and settings determine the conversion factor empirically.

Include negative and positive controls in all experiment. Cells expressing your vector without an insert are a good negative control and you may use well-characterized cells or membranes as your positive control. Such a positive control allows troubleshooting and serves as internal control.

5.8.1.1 Day 1: *E. coli* Transformation

For detailed information about transformation always check specific requirement of various strains with the supplier. For an initial screen, the suggested strains are: BL21 Star (Life technologies #C6010-03), CD43 [24] and BL21 Tuner (Merck Millipore #70622); consider Lemo21 as an alternative (New England Biolabs #C2528H).

Solubilize codon-optimized plasmids from a commercial supplier to a final concentration of 100 ng/μL in milliQ-water. Add 1 μL of DNA to 25 μL of competent cells and leave on ice for 30 min, use a 24 deep-well plate (Whatman, Uniplate #7701-5110) for multiple transformation. Heat shock cells for 2 min in a 42 °C water bath, and place the deep-well plate back on ice for 2 min. Add 225 μL of pre-warmed SOC medium (Life technologies, #15544-034) and incubate in a shaking incubator at 220 rpm and 37 °C for 1 h. Plate 100 μL onto a pre-warmed LB-agar plate containing the appropriate antibiotics and incubate at 37 °C overnight. Note this manual step can be labor intensive when screening large numbers of constructs; with homogeneous high-quality plasmid consider to add 3 mL antibiotics containing LB-medium to the 24 well plate for overnight growth.

5.8.1.2 Day 2: Overnight Culture in Non-inducing Terrific Broth (TB) or 32Y Medium [25]

Fill the appropriate number of wells in 24 deep-well plates with 3 mL of media supplemented with antibiotics. From each transformation plate pick 3–4 colonies and inoculate that them in well with medium. Seal the plate with air-pore seal (Qiagen, #19571) and incubate overnight at 37 °C with a shaking rate of 220 rpm.

5.8.1.3 Day 3: Autoinduction in 1xNPS(64), TB and 32Y Medium

Check by eye that the overnight culture is turbid. For each construct and *E. coli* strain, add 3 mL media prepared with autoinduction cocktail [3 mM $MgCl_2$, 0.8 % (v/v) glycerol, 0.02 % (w/v) glucose] to 2 wells of a fresh 24 deep-well block. Screen inducer concentration using 0 %, 0.02 %, 0.1 % and 0.2 % (w/v) α-lactose (Fig. 5.2) and sample 1xNPS, TB and 32Y medium. Inoculate the fresh media using a multipipette to transfer 30 μL of each overnight culture. Store the overnight cultures at 4 °C, which will be used later to prepare glycerol stocks. Seal the expression plates with air-pore seal and incubate at 37 °C, 220 rpm. After 5–7 h, check that the culture is slightly turbid as a sign of bacterial growth and reduce the temperature to 18 °C for overnight incubation.

5.8.1.4 Day 3: Optional IPTG Induction in 1xNPS, TB and 32Y Media

Check by eye that the overnight culture is turbid. Fill a 24-well deep-well block with 3 mL media supplemented with appropriate antibiotics, pre-warm at 37 °C. Inoculate with 200 µL overnight culture. Grow for 1 h at 37 °C at 220 rpm, reduce temperature to 18 °C for 30 min. Induce expression with a range of inducer concentrations, or in a sequential process start with 0.1 mM IPTG (Isopropyl β-D-1-thiogalactopyranoside, *e.g.*, Life technologies #15529-019).

5.8.1.5 Day 4: Harvest

Cells are harvested ~16–20 h after growth and induction. Spin plates at 2500 × *g*, for 10 min at 4 °C. Carefully discard the supernatant by inverting the plate or pipetting off excess liquid. Discard the spent media with antibiotic according to your local routines and regulations. Continue with lysis and analysis, or freeze the cell pellet in liquid nitrogen and store the plate at −20 °C for later analysis.

5.8.1.6 FSEC Analysis

Typical FSEC results are shown in Fig. 5.3 for MraY ortholog screening in *E. coli* BL21 Star cells. The target expression level determines the amount of cells used for screening. For low expression levels, it is recommended to start from resuspensions of 10 OD (optical density, or absorbance, units at 600 nm) for *E. coli* or, equivalently, from 5 to 10×10^5 cells when mammalian cells are used. These numbers can be reduced for better expressing targets. The optimal buffer systems are protein-dependent and need to be determined experimentally, however ensure that the screening buffers are compatible with the chosen size-exclusion columns.

1. Resuspend all cells to an OD of 10/mL in resuspension buffer; as initial suggestion use 50 mM HEPES pH 7.5, 500 mM NaCl, 20 % glycerol, 1× EDTA-free protease inhibitor cocktail (Roche, #11873580001).
2. *E. coli* cells should be lysed prior to solubilization. For a plate-based approach, resuspend cells in a buffer containing lysozyme (Fluka #62971) at 100 µg/mL, and benzonase (Novagen #70746) 2 at U/ml to reduce sample viscosity from DNA. Subject the plate to three freeze-thaw cycles using liquid nitrogen and a water bath at 42 °C.
3. Distribute solubilization buffer in a 96-well block (Thermo #295-100661). As initial suggestion for solubilization use: 2 % (w/v) n-Dodecyl β-D-maltoside (DDM; CAS 69227-93-6), 0.6 % (w/v) 3-[(3-cholamidopropyl)dimethylammonio]-2-hydroxy-1-propanesulfonate (CHAPS; CAS 75621-03-3) and 0.2 % (w/v) cholesteryl hemisuccinate (CHS; CAS 102601-49-0), dissolved in resuspension buffer.
4. Transfer an equal volume of lysed *E. coli* cells to the solubilization buffer, mix by pipetting. Final sample volume including cells should be at least 200 µL.
5. Solubilize by incubating the plate on an orbital plate shaker (*e.g.*, VARIOMAG mono shaker) with significant stirring for at least 1 h in the cold room.
6. Equilibrate a filter plate (0.2 µm Bioinert membrane, PALL # PN5042) by filtration of 100 µL FSEC buffer [*i.e.*, 50 mM HEPES pH 7.5, 500 mM NaCl, 10 % (v/v) glycerol and 0.05 % (w/v) DDM and 0.005 % (w/v) CHS)]. Secure the filter plate on top of collection plate (Greiner #651201) with ordinary lab tape and centrifuge at 3700× *g* for 5 min in a swing-out rotor suitable for plate centrifugation. After centrifugation secure the filter plate on top of a fresh collection plate.
7. Transfer 200 µL of the solubilization mix to the filter plate and remove insoluble material by centrifuging at 3700× *g* for 15 min, 4 °C. Thirty microliter filtrate is sufficient for FSEC-HPLC analysis.
8. Add the His-specific probe (Fig. 5.4) to each sample to a final concentration of 0.8 µM. The probe synthesis is described in detail in [67]; in brief, 1 mg of peptide labeled with FAM-fluorophore was resuspended in PBS buffer with 2 mM TCEP and incubated for 2 h at room temperature. Ten milligrams of maleimido-C3-NTA was added and incu-

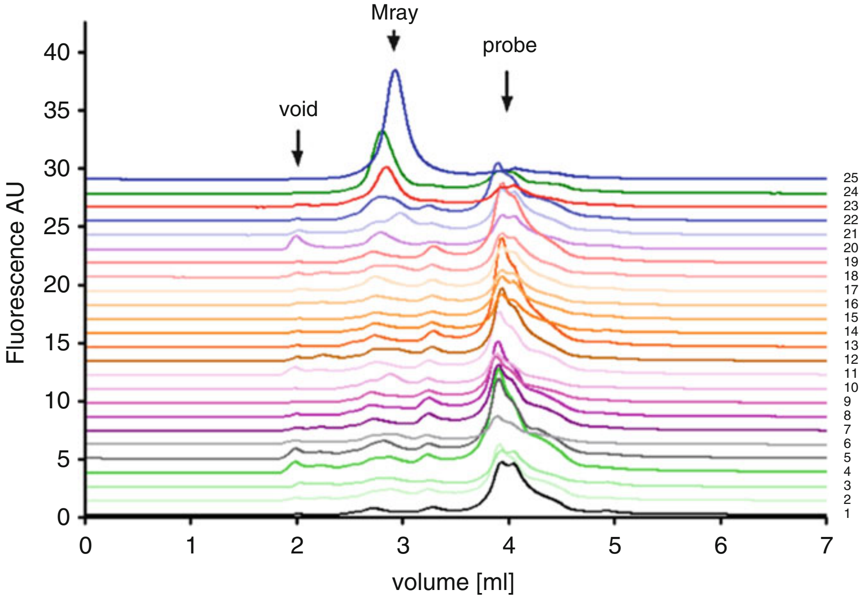


Fig. 5.3 Ortholog screening of a bacterial membrane protein using FSEC. Orthologs were expressed in *E. coli*, and whole cell lysates were solubilized with dodecylmaltoside prior to addition of probe. Monodisperse protein elutes at ~2.9 mL, whereas the free probe elutes at 3.9 mL. Cells transformed with empty expression vector were used as a negative control. To ensure that the full expression range was observed, detergent-solubilized samples were analyzed with increasing probe concentrations until a significant peak of free probe was observed. *1* empty control, *2 C. diphteriae*, *3 C. testosteroni*, *4 C. trachomatis*, *5 E. coli*, *6 E. faecalis*, *7 F. nodosum*, *8 F. placidus*, *9 F. prausnitzii*, *10 H. neptunium*, *11 H. pylori*, *12 M. thermophila*, *13 M. voltae*, *14 P. gingivalis*, *15 P. horikoshii*, *16 R. rickettsii*, *17 S. pneumoniae*, *18 T. mathranii*, *19 T. neutrophilus*, *20 M. thermautotrophicus*, *21 M. thermoacetica*, *22 L. gasicomitatum*, *23 T. thermophilus*, *24 C. bolteae*, *25 B. subtilus* (The figure is reproduced from Ref. [67] with permission from Wiley-Blackwell)

bated for 24 h in the dark. The peptide was purified and loaded with Ni^{2+} on streptactin agarose resin. The probe is stored at 80 μM in small aliquots at −80 °C.

9. Incubate samples on ice for >2 h prior to HPLC analysis. Separate 10–20 μL sample on a HPLC (with a plate autosampler) using size exclusion chromatography, *e.g.*, Agilent BioSec-3 300-Å 4.6 × 300 mm, at 0.3 mL/min for 20 min. Collect fluorescent signal using an excitation wavelength of 482 nm and emission wavelength of 520 nm. Much shorter columns can be run as well albeit with a loss of resolution, *i.e.*, Agilent BioSec-3 300-Å 7.8 × 50 mm at 1.2 mL/min reduces measurements to 2.5 min per sample.

10. The fluorescent probe binds to the His-tag via Ni^{2+}-ion interactions similar to Ni-IMAC purification; therefore detection may be impaired by the salt concentration (150–600 mM in the sample is acceptable), reducing agents (TCEP up to 5 mM is acceptable) and presence of divalent cations. EDTA or high concentrations of imidazole should be avoided. Parallel to FSEC analysis, use immunoblotting to confirm membrane protein expression. Select which constructs to progress and ensure that cryo-stocks are prepared from the overnight growths by addition of, *e.g.*, 7 % (v/v) sterile DMSO.

11. The procedures above can be adapted to screen effects on membrane protein quality of various buffers, detergents, lipid additives or (small molecule) binders.

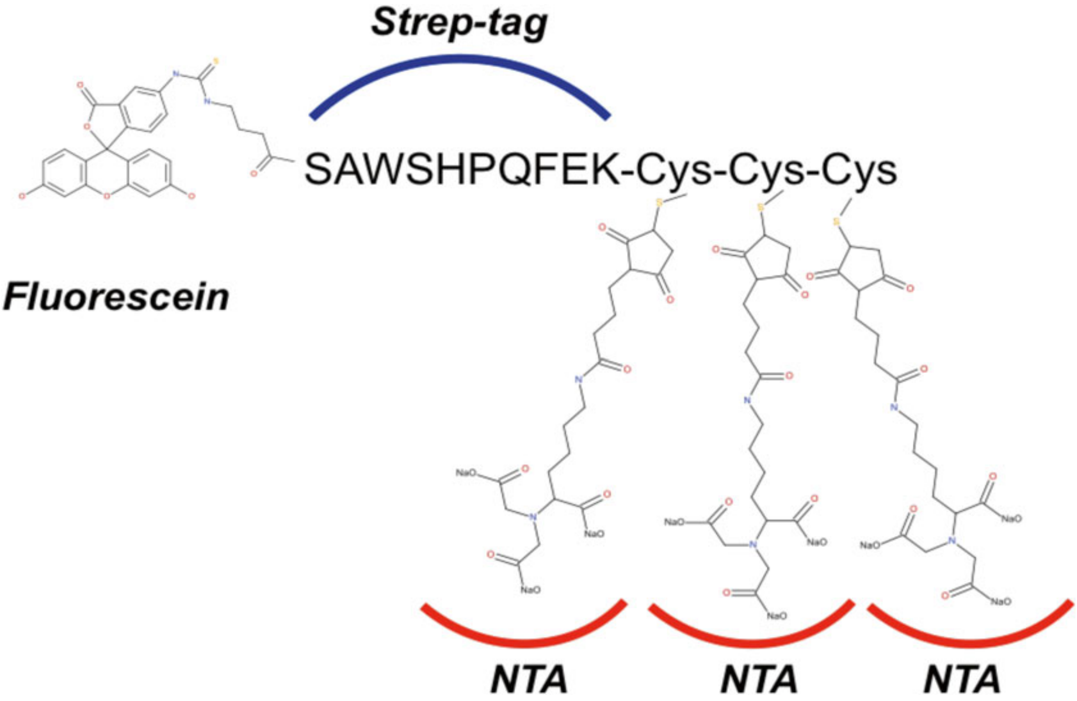


Fig. 5.4 **Structure of the peptide based fluorescent FSEC probe** (The figure is reproduced from Ref. [67] with permission from Wiley-Blackwell)

References

1. Egloff P, Hillenbrand M, Klenk C, Batyuk A, Heine P, Balada S et al (2014) Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in escherichia coli. Proc Natl Acad Sci U S A 111(6):E655–E662
2. Grisshammer R, Duckworth R, Henderson R (1993) Expression of a rat neurotensin receptor in escherichia coli. Biochem J 295(Pt 2):571–6
3. Weiss HM, Grisshammer R (2002) Purification and characterization of the human adenosine A(2a) receptor functionally expressed in escherichia coli. Eur J Biochem 269(1):82–92
4. Thoren S, Weinander R, Saha S, Jegerschold C, Pettersson PL, Samuelsson B et al (2003) Human microsomal prostaglandin E synthase-1: purification, functional characterization, and projection structure determination. J Biol Chem 278(25):22199–22209
5. Xu S, McKeever BM, Wisniewski D, Miller DK, Spencer RH, Chu L et al (2007) Expression, purification and crystallization of human 5-lipoxygenase-activating protein with leukotriene-biosynthesis inhibitors. Acta Crystallogr Sect F: Struct Biol Cryst Commun 63(Pt 12):1054–1057
6. Quick M, Loo DD, Wright EM (2001) Neutralization of a conserved amino acid residue in the human Na+/glucose transporter (hSGLT1) generates a glucose-gated H+ channel. J Biol Chem 276(3):1728–1734
7. Srinivasan V, Pierik AJ, Lill R (2014) Crystal structures of nucleotide-free and glutathione-bound mitochondrial ABC transporter Atm1. Science 343(6175):1137–1140
8. Park SH, Das BB, Casagrande F, Tian Y, Nothnagel HJ, Chu M et al (2012) Structure of the chemokine receptor CXCR1 in phospholipid bilayers. Nature 491(7426):779–783
9. Van den Berg B, Clemons WM Jr, Collinson I, Modis Y, Hartmann E, Harrison SC et al (2004) X-ray structure of a protein-conducting channel. Nature 427(6969):36–44
10. Voorhees RM, Fernandez IS, Scheres SH, Hegde RS (2014) Structure of the mammalian ribosome-Sec61 complex to 3.4 A resolution. Cell 157(7):1632–1643
11. Mitra K, Schaffitzel C, Shaikh T, Tama F, Jenni S, Brooks CL 3rd et al (2005) Structure of the E. coli protein-conducting channel bound to a translating ribosome. Nature 438(7066):318–324
12. Kramer G, Boehringer D, Ban N, Bukau B (2009) The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. Nat Struct Mol Biol 16(6):589–597
13. Driessen AJ, Nouwen N (2008) Protein translocation across the bacterial cytoplasmic membrane. Annu Rev Biochem 77:643–667

14. Pedersen S (1984) Escherichia coli ribosomes translate in vivo with variable rate. EMBO J 3(12):2895–2898
15. Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. J Mol Biol 396(5):1310–1318
16. Li GW, Oh E, Weissman JS (2012) The anti-shine-dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484(7395):538–541
17. Fluman N, Navon S, Bibi E, Pilpel Y (2014) mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. Elife 3. doi:10.7554/eLife.03440
18. Kim HS, Ernst JA, Brown C, Bostrom J, Fuh G, Lee CV et al (2012) Translation levels control multi-spanning membrane protein expression. PLoS ONE 7(4), e35844
19. Raue U, Oellerer S, Rospert S (2007) Association of protein biogenesis factors at the yeast ribosomal tunnel exit is affected by the translational status and nascent polypeptide sequence. J Biol Chem 282(11):7809–7816
20. Klepsch MM, Persson JO, de Gier JW (2011) Consequences of the overexpression of a eukaryotic membrane protein, the human KDEL receptor, in escherichia coli. J Mol Biol 407(4):532–542
21. Wagner S, Klepsch MM, Schlegel S, Appel A, Draheim R, Tarry M et al (2008) Tuning escherichia coli for membrane protein overexpression. Proc Natl Acad Sci U S A 105(38):14371–14376
22. Schlegel S, Lofblom J, Lee C, Hjelm A, Klepsch M, Strous M et al (2012) Optimizing membrane protein overexpression in the escherichia coli strain Lemo21(DE3). J Mol Biol 423(4):648–659
23. Narayanan A, Ridilla M, Yernool DA (2011) Restrained expression, a method to overproduce toxic membrane proteins by exploiting operator-repressor interactions. Protein Sci 20(1):51–61
24. Miroux B, Walker JE (1996) Over-production of proteins in escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. J Mol Biol 260(3):289–298
25. Gordon E, Horsefield R, Swarts HG, de Pont JJ, Neutze R, Snijder A (2008) Effective high-throughput overproduction of membrane proteins in escherichia coli. Protein Expr Purif 62(1):1–8
26. Nannenga BL, Baneyx F (2011) Reprogramming chaperone pathways to improve membrane protein expression in escherichia coli. Protein Sci 20(8):1411–1420
27. Mitra K, Ubarretxena-Belandia I, Taguchi T, Warren G, Engelman DM (2004) Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol. Proc Natl Acad Sci U S A 101(12):4083–4088
28. Pogozheva ID, Tristram-Nagle S, Mosberg HI, Lomize AL (2013) Structural adaptations of proteins to different biological membranes. Biochim Biophys Acta 1828(11):2592–2608
29. Dumas F, Tocanne JF, Leblanc G, Lebrun MC (2000) Consequences of hydrophobic mismatch between lipids and melibiose permease on melibiose transport. Biochemistry 39(16):4846–4854
30. East JM, Melville D, Lee AG (1985) Exchange rates and numbers of annular lipids for the calcium and magnesium ion dependent adenosinetriphosphatase. Biochemistry 24(11):2615–2623
31. Lee AG (2011) Biological membranes: the importance of molecular detail. Trends Biochem Sci 36(9):493–500
32. Lee AG (2003) Lipid–protein interactions in biological membranes: a structural perspective. Biochim Biophys Acta (BBA) Biomembr 1612(1):1–40
33. Laganowsky A, Reading E, Allison TM, Ulmschneider MB, Degiacomi MT, Baldwin AJ et al (2014) Membrane proteins bind lipids selectively to modulate their structure and function. Nature 510(7503):172–175
34. Sultan A, Luo M, Yu Q, Riederer B, Xia W, Chen M et al (2013) Differential association of the Na+/H+ exchanger regulatory factor (NHERF) family of adaptor proteins with the raft- and the non-raft brush border membrane fractions of NHE3. Cell Physiol Biochem 32(5):1386–1402
35. Dowhan W, Bogdanov M (2011) Lipid-protein interactions as determinants of membrane protein structure and function. Biochem Soc Trans 39(3):767–774
36. Valiyaveetil FI, Zhou Y, MacKinnon R (2002) Lipids in the structure, folding, and function of the KcsA K+ channel. Biochemistry 41(35):10771–10777
37. Nury H, Dahout-Gonzalez C, Trezeguet V, Lauquin G, Brandolin G, Pebay-Peyroula E (2005) Structural basis for lipid-mediated interactions between mitochondrial ADP/ATP carrier monomers. FEBS Lett 579(27):6031–6036
38. Palsdottir H, Hunte C (2004) Lipids in membrane protein structures. Biochim Biophys Acta 1666(1–2):2–18
39. Hanson MA, Cherezov V, Griffith MT, Roth CB, Jaakola V, Chien EYT et al (2008) A specific cholesterol binding site is established by the 2.8 Å structure of the human β2-adrenergic receptor. Structure 16(6):897–905
40. Studier FW, Moffatt BA (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189(1):113–130
41. Studier FW (1991) Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. J Mol Biol 219(1):37–44
42. Krepkiy D, Wong K, Gawrisch K, Yeliseev A (2006) Bacterial expression of functional, biotinylated peripheral cannabinoid receptor CB2. Protein Expr Purif 49(1):60–70
43. Arechaga I, Miroux B, Karrasch S, Huijbregts R, de Kruijff B, Runswick MJ et al (2000) Characterisation

of new intracellular membranes in escherichia coli accompanying large scale over-production of the b subunit of F(1)F(o) ATP synthase. FEBS Lett 482(3):215–219

44. Massey-Gendel E, Zhao A, Boulting G, Kim HY, Balamotis MA, Seligman LM et al (2009) Genetic selection system for improving recombinant membrane protein expression in E. coli. Protein Sci 18(2):372–383
45. Link AJ, Skretas G, Strauch EM, Chari NS, Georgiou G (2008) Efficient production of membrane-integrated and detergent-soluble G protein-coupled receptors in escherichia coli. Protein Sci 17(10):1857–1863
46. Norholm MH, Light S, Virkki MT, Elofsson A, von Heijne G, Daley DO (2012) Manipulating the genetic code for membrane protein production: what have we learnt so far? Biochim Biophys Acta 1818(4):1091–1096
47. Norholm MH, Toddo S, Virkki MT, Light S, von Heijne G, Daley DO (2013) Improved production of membrane proteins in escherichia coli by selective codon substitutions. FEBS Lett 587(15):2352–2358
48. Bondarenko V, Mowrey DD, Tillman TS, Seyoum E, Xu Y, Tang P (2014) NMR structures of the human alpha7 nAChR transmembrane domain and associated anesthetic binding sites. Biochim Biophys Acta 1838(5):1389–1395
49. Lau FW, Nauli S, Zhou Y, Bowie JU (1999) Changing single side-chains can greatly enhance the resistance of a membrane protein to irreversible inactivation. J Mol Biol 290(2):559–564
50. Warne T, Serrano-Vega MJ, Tate CG, Schertler GF (2009) Development and crystallization of a minimal thermostabilised G protein-coupled receptor. Protein Expr Purif 65(2):204–213
51. Scott DJ, Pluckthun A (2013) Direct molecular evolution of detergent-stable G protein-coupled receptors using polymer encapsulated cells. J Mol Biol 425(3):662–677
52. Molina DM, Cornvik T, Eshaghi S, Haeggström J, Nordlund P, Sabet MI (2008) Engineering membrane protein overproduction in *escherichia coli*. Protein Sci: Publ Protein Soc 17(4):673–680
53. Schlinkmann KM, Hillenbrand M, Rittner A, Kunz M, Strohner R, Pluckthun A (2012) Maximizing detergent stability and functional expression of a GPCR by exhaustive recombination and evolution. J Mol Biol 422(3):414–428
54. Chen KY, Zhou F, Fryszczyn BG, Barth P (2012) Naturally evolved G protein-coupled receptors adopt metastable conformations. Proc Natl Acad Sci U S A 109(33):13284–13289
55. Chun E, Thompson AA, Liu W, Roth CB, Griffith MT, Katritch V et al (2012) Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. Structure 20(6):967–976
56. Tucker J, Grisshammer R (1996) Purification of a rat neurotensin receptor expressed in escherichia coli. Biochem J 317(Pt 3):891–9
57. Mohanty AK, Wiener MC (2004) Membrane protein expression and production: effects of polyhistidine tag length and position. Protein Expr Purif 33(2):311–325
58. Low C, Moberg P, Quistgaard EM, Hedren M, Guettou F, Frauenfeld J et al (2013) High-throughput analytical gel filtration screening of integral membrane proteins for structural studies. Biochim Biophys Acta 1830(6):3497–3508
59. Roosild TP, Greenwald J, Vega M, Castronovo S, Riek R, Choe S (2005) NMR structure of mistic, a membrane-integrating protein for membrane protein expression. Science 307(5713):1317–1321
60. Chowdhury A, Feng R, Tong Q, Zhang Y, Xie XQ (2012) Mistic and TarCF as fusion protein partners for functional expression of the cannabinoid receptor 2 in escherichia coli. Protein Expr Purif 83(2):128–134
61. Deniaud A, Bernaudat F, Frelet-Barrand A, Juillan-Binard C, Vernet T, Rolland N et al (2011) Expression of a chloroplast ATP/ADP transporter in E. coli membranes: behind the mistic strategy. Biochim Biophys Acta 1808(8):2059–2066
62. Neophytou I, Harvey R, Lawrence J, Marsh P, Panaretou B, Barlow D (2007) Eukaryotic integral membrane protein expression utilizing the escherichia coli glycerol-conducting channel protein (GlpF). Appl Microbiol Biotechnol 77(2):375–381
63. Studier FW (2005) Protein production by auto-induction in high density shaking cultures. Protein Expr Purif 41(1):207–234
64. Barria C, Malecki M, Arraiano CM (2013) Bacterial adaptation to cold. Microbiology 159(Pt 12):2437–2443
65. Lundstrom K, Wagner R, Reinhart C, Desmyter A, Cherouati N, Magnin T et al (2006) Structural genomics on membrane proteins: comparison of more than 100 GPCRs in 3 expression systems. J Struct Funct Genomics 7(2):77–91
66. Kawate T, Gouaux E (2006) Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. Structure 14(4):673–681
67. Backmark AE, Olivier N, Snijder A, Gordon E, Dekker N, Ferguson AD (2013) Fluorescent probe for high-throughput screening of membrane protein expression. Protein Sci 22(8):1124–1132
68. Stein A, Weber G, Wahl MC, Jahn R (2009) Helical extension of the neuronal SNARE complex into the membrane. Nature 460(7254):525–528
69. Bondarenko V, Mowrey D, Tillman T, Cui T, Liu LT, Xu Y et al (2012) NMR structures of the transmembrane domains of the alpha4beta2 nAChR. Biochim Biophys Acta 1818(5):1261–1268
70. Mulligan C, Leech AP, Kelly DJ, Thomas GH (2012) The membrane proteins SiaQ and SiaM form an essential stoichiometric complex in the sialic acid tripartite ATP-independent periplasmic (TRAP) transporter SiaPQM (VC1777-1779) from vibrio cholerae. J Biol Chem 287(5):3598–3608

71. Deckers-Hebestreit G (2013) Assembly of the escherichia coli FoF1 ATP synthase involves distinct subcomplex formation. Biochem Soc Trans 41(5):1288–1293
72. von Meyenburg K, Jorgensen BB, Michelsen O, Sorensen L, McCarthy JE (1985) Proton conduction by subunit a of the membrane-bound ATP synthase of escherichia coli revealed after induced overproduction. EMBO J 4(9):2357–2363
73. McCarthy JE, Bokelmann C (1988) Determinants of translational initiation efficiency in the atp operon of escherichia coli. Mol Microbiol 2(4):455–465
74. Patel AM, Dunn SD (1995) Degradation of escherichia coli uncB mRNA by multiple endonucleolytic cleavages. J Bacteriol 177(14):3917–3922
75. Arechaga I, Miroux B, Runswick MJ, Walker JE (2003) Over-expression of escherichia coli F1F(o)-ATPase subunit a is inhibited by instability of the uncB gene transcript. FEBS Lett 547(1–3):97–100
76. Arechaga I, Butler PJ, Walker JE (2002) Self-assembly of ATP synthase subunit c rings. FEBS Lett 515(1–3):189–193
77. Hermolin J, Fillingame RH (1995) Assembly of F0 sector of escherichia coli H+ ATP synthase. interdependence of subunit insertion into the membrane. J Biol Chem 270(6):2815–2817
78. Pierson HE, Uhlemann EM, Dmitriev OY (2011) Interaction with monomeric subunit c drives insertion of ATP synthase subunit a into the membrane and primes a-c complex formation. J Biol Chem 286(44):38583–38591
79. Lata S, Gavutis M, Tampe R, Piehler J (2006) Specific and stable fluorescence labeling of histidine-tagged proteins for dissecting multi-protein complex formation. J Am Chem Soc 128(7):2365–2372
80. Jaremko L, Jaremko M, Giller K, Becker S, Zweckstetter M (2014) Structure of the mitochondrial translocator protein in complex with a diagnostic ligand. Science 343(6177):1363–1366
81. Yang J, Ma YQ, Page RC, Misra S, Plow EF, Qin J (2009) Structure of an integrin alphaIIb beta3 transmembrane-cytoplasmic heterocomplex provides insight into integrin activation. Proc Natl Acad Sci U S A 106(42):17729–17734
82. Mowrey DD, Cui T, Jia Y, Ma D, Makhov AM, Zhang P et al (2013) Open-channel structures of the human glycine receptor alpha1 full-length transmembrane domain. Structure 21(10):1897–1904
83. Berardi MJ, Shih WM, Harrison SC, Chou JJ (2011) Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. Nature 476(7358):109–113
84. Teriete P, Franzin CM, Choi J, Marassi FM (2007) Structure of the na, K-ATPase regulatory protein FXYD1 in micelles. Biochemistry 46(23):6774–6783
85. Oxenoid K, Chou JJ (2005) The structure of phospholamban pentamer reveals a channel-like architecture in membranes. Proc Natl Acad Sci U S A 102(31):10870–10875
86. Ellena JF, Liang B, Wiktor M, Stein A, Cafiso DS, Jahn R et al (2009) Dynamic structure of lipid-bound synaptobrevin suggests a nucleation-propagation mechanism for trans-SNARE complex formation. Proc Natl Acad Sci U S A 106(48):20306–20311
87. Ferguson AD, McKeever BM, Xu S, Wisniewski D, Miller DK, Yamin TT et al (2007) Crystal structure of inhibitor-bound human 5-lipoxygenase-activating protein. Science 317(5837):510–512
88. Lu P, Ma D, Yan C, Gong X, Du M, Shi Y (2014) Structure and mechanism of a eukaryotic transmembrane ascorbate-dependent oxidoreductase. Proc Natl Acad Sci U S A 111(5):1813–1818
89. Newby ZE, O'Connell J 3rd, Robles-Colmenares Y, Khademi S, Miercke LJ, Stroud RM (2008) Crystal structure of the aquaglyceroporin PfAQP from the malarial parasite plasmodium falciparum. Nat Struct Mol Biol 15(6):619–625
90. Nishida M, Cadene M, Chait BT, MacKinnon R (2007) Crystal structure of a Kir3.1-prokaryotic kir channel chimera. EMBO J 26(17):4005–4015

6 Cell-Free Synthesis of Macromolecular Complexes

Mathieu Botte*, Aurélien Deniaud*,
and Christiane Schaffitzel

Abstract

Cell-free protein synthesis based on *E. coli* cell extracts has been described for the first time more than 50 years ago. To date, cell-free synthesis is widely used for the preparation of toxic proteins, for studies of the translation process and its regulation as well as for the incorporation of artificial or labeled amino acids into a polypeptide chain. Many efforts have been directed towards establishing cell-free expression as a standard method for gene expression, with limited success. In this chapter we will describe the state-of-the-art of cell-free expression, extract preparation methods and recent examples for successful applications of cell-free synthesis of macromolecular complexes.

Keywords

Cell-free synthesis • Ribosomes • Membrane proteins • Synthetic biology

*Author contributed equally with all other contributors.

M. Botte
European Molecular Biology Laboratory, Grenoble
Outstation, 71 avenue des Martyrs, 38042 Grenoble
Cedex 9, France

A. Deniaud (✉)
CEA Grenoble, iRTSV/LCBM/BioMet,
17 rue des Martyrs, 38054 Grenoble, France

European Molecular Biology Laboratory, Grenoble
Outstation, 71 avenue des Martyrs, 38042 Grenoble
Cedex 9, France
e-mail: aurelien.deniaud@cea.fr

C. Schaffitzel (✉)
European Molecular Biology Laboratory, Grenoble
Outstation, 71 avenue des Martyrs, 38042 Grenoble
Cedex 9, France

School of Biochemistry, University of Bristol,
Bristol BS8 1TD, UK
e-mail: schaffitzel@embl.fr

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_6

6.1 Introduction: Motivation and Challenges

The capacity of cell extracts to synthesize proteins has been shown in the 1950s of the last century [1, 2], several years before the identification of ribosomes as protein-synthesizing machines [3]. The cell-free extract was based on the classical S30 fraction obtained by a 30,000× g centrifugation step at 4 °C for 1 h. Initially, endogenous mRNA was used for *in vitro* translation [4]. Subsequently, Nirenberg and Matthaei developed a protocol to degrade endogenous messenger RNA present in the cell extract and to add exogenous mRNA [5, 6]. The first cell-free protein synthesis (CFPS) from DNA, using a so-called coupled transcription-translation system was developed in the late 1960s by the group of Zubay [7]. They used their coupled transcription-translation system to study the regulation of gene expression by the *E. coli* lactose operon. Most cell-free extract preparation and *in vitro* translation protocols are based on this protocol [8, 9].

Significant improvements with respect to protein yields were achieved in the late 1980s, in particular by the group of Spirin, which established the use of phage-specific RNA polymerases, SP6 [10] or T7 RNA polymerases [8]. Using these polymerases a high level of a specific mRNA during the *in vitro* transcription-translation reaction can be achieved and maintained. Importantly, the Spirin laboratory described the first 'continuous' *in vitro* translation system. It allows for a continuous exchange of small molecules between a 'feeding compartment' providing energy and substrates (amino acids) for the translation reaction and a 'reaction compartment' from which inhibitory reaction products are removed by dialysis [10, 11]. In a continuous set-up, the *in vitro* translation reaction can continue for several hours or even days, compared to 40–60 min using the classical reaction set-up. This allows obtaining significantly increased yields: for instance 6 mg chloramphenicol acetyl-transferase protein per milliliter of *in vitro* translation reaction were synthesized in 21 h [12].

With these advancements, cell-free translation became a very interesting technology for protein production in structural biology. In particular the RIKEN Structural Genomics/Proteomics Initiative (RSGI) in Japan invested into the automation of cell-free protein synthesis and high-throughput screening of protein products with the aim to obtain high yields of isotope-labeled proteins for NMR studies [13–15]. Notably, specific ^{15}N and ^{13}C labeling for any amino acid is trivial as soon as the protein is expressed *in vitro*. Accordingly, numerous NMR structures have been solved using this approach [16–18]. CFPS also led to several X-ray structures [15, 19].

CFPS allows the rapid and economical screening of a number of different proteins or protein variants (mutants, truncations, etc.) when these are required only in small quantities. Classical sub-cloning of constructs into plasmids is not required since the *in vitro* transcription/translation reaction can be started from PCR products [20] which significantly improves the screening capacities, a high-throughput set-up and automation.

Cell-free expression remains a powerful approach for the production of toxic and insoluble proteins, for instance membrane proteins. The group of F. Bernhard significantly improved *in vitro* translation protocols to be able to produce membrane proteins in the presence of detergents or of lipids (for review [21]). Subsequent crystallization attempts, for instance of G-protein coupled receptors, remained mostly unsuccessful. To date, three membrane proteins which were produced by *in vitro* translation have been crystallized: VDAC, diacylglycerol kinase and EmrE [22–24]. In the case of EmrE, cell-free synthesis was used to generate a seleno-methionine derivative in order to phase already existing crystallographic data.

Cell-free expression has thus several very attractive applications, related to the expression of toxic proteins, rapid production of small quantities of proteins for screening and protein engineering purposes as well as for the incorporation of unnatural amino acids in structural and synthetic biology. In this chapter, we describe the different *in vitro* translation reaction set-ups used in the field, and we present successful applications applied to study large macromolecular complexes.

6.2 Basics of *E. coli* Transcription/Translation Systems

6.2.1 The Classical S30-Based Cell-Free Expression System

The most common method for cell-free expression is using *E. coli* S30 extract [7]. This classical cell-free expression system has been only slightly modified since the first description of the protocol by Nirenberg [5]. The S30 extract is composed of a soluble fraction which is obtained after lysis of *E. coli* cells and centrifugation of the lysate at 30,000× g. Thus, this extract contains all the cytosolic enzymes required for transcription and translation. However, without further treatment, the extract contains endogenous mRNAs which will also be translated, leading to unwanted side products. Nirenberg and Matthaei established a protocol to remove endogenous mRNAs without destabilizing the ribosomal RNAs [6]: After centrifugation, the lysate is treated with high-salt concentrations resulting in the release of mRNAs from the ribosomes. The endogenous mRNAs is then degraded by the RNases present in the cell extract (*e.g.*, by incubation for 1 h at 25 °C). For the cell-free transcription/translation reaction, the S30 extract needs to be supplemented with the 20 amino acids, the *E. coli* tRNAs, nucleotides (ATP, GTP, CTP and UTP) required as energy sources as well as building blocks for the RNA synthesis, as well as an energy-regenerating system composed of phosphoenol pyruvate and pyruvate kinase, and the T7 RNA polymerase for efficient *in vitro* transcription [9]. Alternative energy regeneration systems have been reported such as acetyl kinase and acetyl phosphate or creatine phosphate and creatine kinase [12, 25].

The template used in the classical cell-free expression system can be either plasmid DNA, linear DNA (PCR products) or mRNA. Usually, *in vitro* translated proteins are tagged for subsequent affinity purification directly from the cell-free expression reaction. In addition to the basic components, co-factors or regulatory proteins which are not present or under-represented in the *E. coli* S30 extract can be added for the production of specific proteins. For instance, Yang and Zubay showed that the *ara*C protein which is required for gene expression of the *ara* operon was lost during the S30 extract preparation, thus inhibiting the expression of proteins from the *ara* operon [26]. The addition of chaperones such as Trigger Factor, DnaK, DnaJ, GrpE, GroEL/GroES and protein disulfide isomerase often increases the amount of soluble proteins and helps folding of disulfide-containing proteins (*e.g.* immunoglobulin domains, see applications) [27]. In conclusion, the composition of the classical cell-free expression system can be optimized and tailored according to the specific requirements of the expressed protein.

The yields (between few micrograms up to several milligrams per milliliter reaction) are dependent on the expressed protein, its mRNA stability, the composition of the cell-free reaction mixture and on the experimental set-up. For cell-free expression, two configurations can be used. The first configuration, which is easier to implement, is the batch method. The bottleneck of this set-up is the yield, which is quite low due to the consumption of energy and amino acids as well as to the accumulation of by-products, which have an inhibitory effect on the *in vitro* transcription/translation reaction. As a consequence, using the batch method protein is produced mostly during the first 60 min of *in vitro* translation, thus limiting the yield.

The second configuration was developed to overcome this problem: the continuous exchange cell-free (CECF) system [11] (Fig. 6.1). This system is divided in two compartments that can exchange low molecular weight compounds through a dialysis membrane. The reaction compartment contains all the high molecular weight species required for the reaction such as the cell extract, the enzymes and the nucleic acids as well as the low molecular weight substrates required for the reaction. The feeding compartment contains only the low molecular weight compounds, *i.e.*, the NTPs, substrates of the energy regeneration system and the amino acids. Usually, the feeding compartment is more than ten-times larger than the reaction compartment. Consequently, during the cell-free expression reaction, which is subjected to mixing or shaking, the by-products are dialyzed from the reaction mixture into the feeding compartment. At the same time, the NTPs, energy substrates and amino acids in the reaction mixture are constantly replenished in the reaction compartment. Using this technique, the cell-free expression reaction can be maintained for tens of hours yielding more than 10 mg of protein per milliliter of reaction [28].

6.2.2 The PURE Cell-Free Expression System

The PURE (Protein synthesis Using Recombinant Elements) system has been developed with the idea to use only purified components of the transcription and translation machinery for *in vitro* synthesis [29]. To this end, initiation, elongation and termination factors as well as the 20 aminoacyl-tRNA synthetases, the methyl-tRNA transformylase and the T7 RNA polymerase were expressed as recombinant proteins with a hexahistidine-tag and affinity purified. In total 31 proteins are added to reconstitute the *in vitro* transcription/translation reaction. The ribosomal subunits were purified from *E. coli* cells and added to the translation reaction. The resulting PURE system can produce about 100 µg of model proteins per ml reaction in one hour (GFP and DHFR). In addition, the PURE system contains

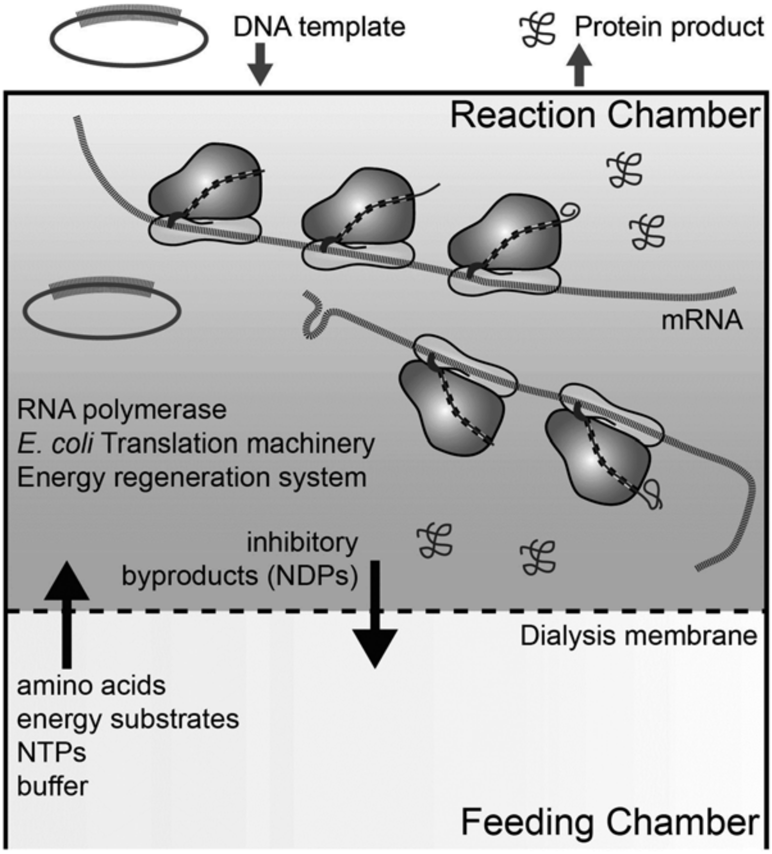


Fig. 6.1 Scheme of the continuous exchange cell-free (CECF) system. Two compartments exist separated by a dialysis membrane: the reaction compartment contains the cell extract with the translation machinery, the template (DNA), the RNA polymerase and the low molecular weight substrates required for *in vitro* transcription and translation. The feeding chamber contains NTPs, substrates of the energy regeneration system and the amino acids in the same reaction buffer as used in the reaction chamber. The feeding chamber is usually more than ten-times larger than the reaction chamber. During protein synthesis, inhibitory side products of the transcription/translation reaction can diffuse into the feeding chamber and thus are diluted. Substrates are consumed during the reaction and are restocked from the feeding compartment

46 tRNAs, NTPs, creatine phosphate, 10-formyl-5,6,7,8-tetrahydrofolic acid, 20 amino acids, creatine kinase, myokinase, nucleoside-diphosphate kinase and pyrophosphatase. Chaperones, heat shock proteins and other factors can be added to the reaction mixture to keep proteins soluble and assist in protein folding.

The use of histidine-tagged translation components offers the possibility to produce the protein of interest without any tag for affinity purification. The newly synthesized protein can still be easily purified in two steps: ultrafiltration to remove the ribosomes and a Ni-NTA affinity chromatography step to remove the recombinant, his-tagged translation factors. Importantly, RNases and proteases are not present in the PURE system. Thus, mRNA of limited stability can still be a template for translation, and proteins which are rapidly degraded *in vivo* can be produced *in vitro*.

The PURE system allows the efficient production of proteins with artificial amino acids. To this end, release factor 1 is omitted from the reaction and a chemically synthesized mis-acylated amino-acyl-tRNA specific for UGA (amber stop codon) is added. This is particularly useful to incorporate fluorescent dyes or specific cross-linkers in the proteins for instance with the aim to analyze protein-protein interactions. Recent improvements of the system aimed at the *in vitro* synthesis of membrane proteins in the presence of lipids. In summary, the PURE system is highly versatile. It can be modified as specific proteins and other factors can be omitted or added to the reaction according to the needs of the proteins to be produced.

6.3 Considerations for Cell-Free Protein Synthesis Experiments and Challenges to Produce Protein Complexes

As outlined above, two major approaches exist for cell-free protein expression using the *E. coli* transcription/translation machinery. The S30 cell extract-based and PURE cell-free systems differ significantly by the degree of purification of the components used. The PURE system has the advantage of being protease- and nuclease-free compared to the S30 cell extract where all cytosolic components are present in the extract. Thus, linear nucleic acids (PCR products and mRNAs) are more stable in the PURE reaction system. Also, proteolytic cleavage of the synthesized protein can be avoided using the PURE system. An additional advantage of the PURE system is the absence of ATP-consuming proteins, which are responsible for the rapid energy depletion in the S30-based system [29]. However, because the PURE system is based on purified components, it is conceivable that some important cofactors or chaperones are missing in this purified system, leading to inefficient folding of the protein. Addition of Trigger Factor, DnaJ, DnaK and GrpE as well as GroEL/GroES may help to improve the yield of soluble, functional protein [30].

The cell extract-based system has the advantage that it is possible to produce the cell extract in large amounts in a standard molecular biology laboratory in a relatively short time (2–3 days for cell extract production and testing). This can be cost-saving, and it allows for upscaling of the *in vitro* translation reaction. The disadvantage of such cell-extract preparations is that batch-dependent differences in translation activity need to be taken into account. This limits the reproducibility of the method.

The expression of multi-protein complexes is challenging *in vitro* and *in vivo*. The correct stoichiometry is difficult to achieve, and the least expressed protein subunit of the complex determines the overall yield of complex. The cell-free systems can be used to express protein complexes: Several DNA templates encoding the protein subunits of the complex can be added simultaneously to the cell-free reaction. In this context, the main advantage of the cell-free expression system compared to the cell-based system is the possibility to precisely adjust the expression of the different subunits of the complex by optimizing the amounts and the ratio of the DNA templates added to the translation reaction. Initial small-scale trials are usually used to optimize the production of the protein subunits in order to achieve stoichiometric expression and homogenous complex formation. In contrast to a cell-based expression in which the different subunits are mostly expressed at the same time, cell-free expression systems allows the sequential addition of DNA templates to the reaction mixture. Moreover, chaperones and additives can be added to the reaction mixture for the efficient integration of the subunits in the complex.

E. coli membrane proteins are mostly dependent on the presence of the conserved Sec translocation machinery for their proper integration into the membrane bilayer [31]. Traditionally, microsomal membranes from dog pancreas treated with high salt and partial trypsin digestion were added to the *in vitro* translation reaction to achieve co-translational protein translocation [32]. *E. coli* inverted membrane vesicles and proteoliposomes reconstituted from components of the translocation machinery also have been successfully used for protein translocation and secretion [33, 34]. However, the specific requirements of membrane proteins for efficient translocation and folding are still poorly understood. Cell-free systems are commonly used to study the process of co-translational membrane protein insertion and folding which is rather inefficient [33]. The presence of membrane protein chaperones and additional translocation factors may be crucial. For instance, the subunit c of the F_1F_0-ATPase has been shown to be dependent on the function of YidC, which is an insertase, integrating small membrane proteins into the membrane of *E. coli* [35]. In Sect. 6.4.2, we describe the application of cell-free expression systems for membrane protein synthesis and their integration into a lipid bilayer or detergent micelles.

6.4 Applications

6.4.1 Ribosome-Nascent Chain Complexes

One important application of cell-free synthesis relates to the preparation of ribosome-nascent chain complexes (RNCs) for structural and functional studies as well as for synthetic biology applications including protein engineering, selection and evolution. To this end, mRNA-ribosome-nascent polypeptide complexes are produced which are stalled in a specific translational state. For this application, cell extract is required containing high concentrations of active ribosomes. The aim is not high yields of newly synthesized protein, but every ribosome is supposed to translate a mRNA template once and then get stalled before translation termination. For RNC production, the *in vitro* transcription and translation reaction are often uncoupled [36]. In a first step, the mRNA template is generated by *in vitro* transcription, using T7 RNA polymerase for instance. The mRNA is subsequently purified by LiCl precipitation followed by ethanol precipitation and added to the *in vitro* translation reaction. The purified mRNA is then added to the *in vitro* translation reaction. The translation reaction can be stopped by addition of high concentrations of magnesium, chloramphenicol or other antibiotics. Alternatively, stalling motifs like SecM or TnaC, or mRNA templates without a stop codon are used to arrest translation. To stabilize the RNCs in the *in vitro* translation reaction mix, it is recommended to add oligonucleotides that inhibit the transfer-messenger-RNA complex of *E. coli* which recognizes ribosomes stalled during protein translation [37]. Subsequently, RNCs can be purified by traditional sucrose gradient centrifugation, or affinity purification via the nascent polypeptide which contains a specific purification tag or an epitope recognized by an antibody [36].

Homogenous RNCs stalled in a specific translational state and complexes with translation factors or factors in co-translational events are mostly studied by single-particle electron cryo-microscopy (cryo-EM). Thanks to recent advances in single-particle cryo-EM it is now possible to reach near-atomic resolution. For instance, a translating ribosome stalled with a TnaC-motif acting L-tryptophan sensor was recently reported at 3.8 Å resolution [38]. Stalling motifs like TnaC and SecM are short peptide sequences that interact with the ribosomal tunnel during translation and induce a conformation in the peptidyl transferase center of the large ribosomal subunits that inhibits further elongation of the nascent polypeptide chain. The ribosome is thus trapped in a specific conformation and displays a nascent polypeptide of defined length. The stalling peptide is hidden in the ribosomal tunnel while the N-terminal part of the nascent chain can exit from the ribosomal tunnel (Fig. 6.2). At the exit of the ribosomal tunnel the nascent chain can fold into a functional protein or bind diverse protein factors involved in co-translational folding, targeting and translocation.

Such factors can be directly added to the cell-free expression system. This was the case with the trigger factor (TF), for instance, which is the first chaperone interacting with the newly synthesized polypeptide exiting the ribosome tunnel [39]. Structure determination of the RNC-TF complex suggested that the co-translational folding of the nascent chain was favored by a protected environment formed by TF and the ribosome (Fig. 6.3). Using a similar approach, several ribosomal complexes have been solved by cryo-EM providing important insights into the molecular mechanism of co-translational targeting and translocation [31]. For these studies a DNA sequence encoding the N-terminal part which includes the signal-anchor sequence of the *E. coli* membrane protein FtsQ was used to produce RNCs. Subsequently, ribosomal complexes were reconstituted for cryo-EM studies by adding purified signal recognition particle (SRP) [40] or SRP-SRP receptor complexes [41] to the RNCs (Fig. 6.3).

Structural insights into the mechanism of signal sequence surveillance during protein targeting were obtained by using a DNA sequence for the *E. coli* autotransporter EspP for RNC generation [42]. EspP is not targeted by SRP to the membrane, but its signal sequence can be bound by SRP. The cryo-EM structures of the RNC-SRP-SRP receptor complexes with the EspP nascent chain revealed how RNCs can be rejected

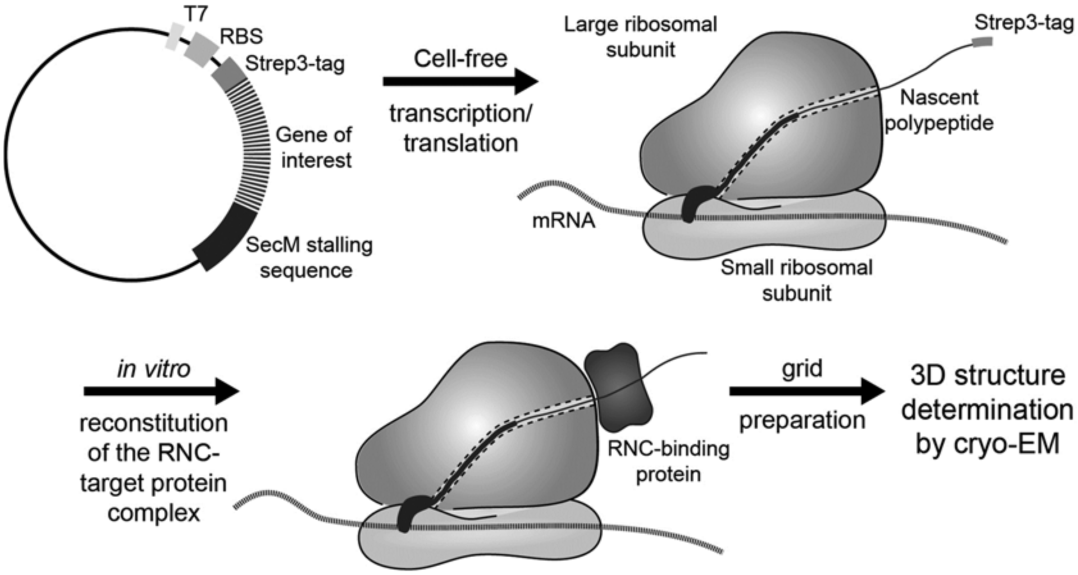


Fig. 6.2 ***In vitro* preparation of ribosome-nascent chain complexes**. The DNA template used encodes a promoter (T7 if T7 RNA polymerase is used for *in vitro* transcription), a Shine-Dalgarno sequence (ribosome binding site), an N-terminal triple TAG (Strep3-tag) followed by the sequence encoding the gene of interest. At the 3′ end the gene encoding the protein of interest is fused in frame to a sequence encoding the translation arrest motif of SecM. During *in vitro* translation, the protein synthesis is not terminated at a stop codon. It is stalled due to the presence of the SecM arrest motif. This results in stable ternary complexes consisting of mRNA, ribosome and nascent polypeptide. The RNCs can be purified via sucrose gradient centrifugation and via the N-terminal tag of the nascent polypeptide by affinity chromatography. Finally, RNCs and binding factors are reconstituted and analysed, for instance by single particle cryo-EM

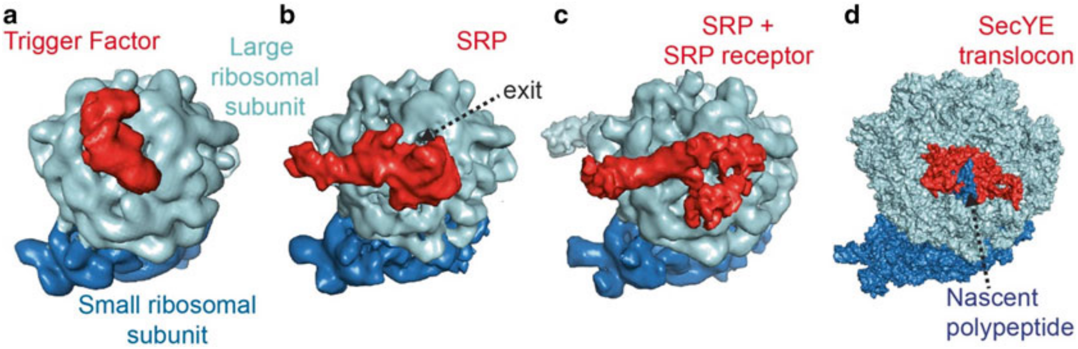


Fig. 6.3 Cryo-EM reconstructions of *E. coli* ribosomal complexes in co-translational folding, targeting and translocation. Homogeneous RNC preparations are used to reconstitute complexes with ribosome binding partners. These complexes allowed visualizing how trigger factor binds to the large ribosomal subunit (50S) and arches over the exit of the ribosomal tunnel (**a**). Together, the ribosome and trigger factor provide a protected folding space for the ribosome [39]. (**b**) The signal recognition particle (SRP) binds next to the exit of ribosomal tunnel and adopts an elongated conformation stabilized by interactions with 50S [40]. (**c**) SRP receptor binding leads to formation of an *early* complex which adopts a V-shape [42]. (**d**) After successful handover of the translating ribosome, the SecYEG complex binds tightly to the exit of the ribosomal tunnel. The translocation channel is aligned with the ribosomal tunnel such that an almost continuous channel from the PTC into the periplasm is formed for the nascent chain [44]. The scheme also visualizes the increasing resolution that can be achieved by single particle cryo-EM due to significant improvements in the microscope, detectors and image processing

from the SRP targeting pathway [42] (Fig. 6.3). These RNCs were crucial to elucidate the conformational states of SRP and its receptor during co-translational targeting by Fluorescence Resonance Energy Transfer (FRET) [43]. These studies revealed that the targeting reaction is tightly controlled in space and time by the ribosome, the translocation machinery and through GTP hydrolysis.

RNCs displaying a signal-anchor sequence have also been successfully used to reconstitute complexes with the *E. coli* protein-conducting channel SecYEG and to solve the structure by cryo-EM [44] (Fig. 6.3). Similarly, RNCs that translate the subunit c of the ATP synthase allowed the reconstitution of complexes with the *E. coli* YidC translocase for cryo-EM [45] and biochemical characterization of the complex using cross-linking agents [35]. In summary, homogenous RNCs are a prerequisite for structural studies. To date, cell-free translation followed by RNC purification and reconstitution of complexes is the method of choice for cryo-EM studies of ribosomal complexes in translation initiation, elongation, termination, recycling and many other ribosomal complexes.

Notably, RNCs are also successfully used to study co-translational folding, targeting and translocation [46]. The dynamic folding of the nascent chain can be studied by FRET and NMR [47, 48]. For NMR, two advantages of the cell-free translation system can be exploited: The specific isotope-labeling of the nascent polypeptide during cell-free synthesis while the ribosome is not labeled, as well as the arrest of the translation reaction to produce nascent chains of different lengths. Moreover, the impact of Trigger Factor and other chaperones on the folding of the nascent chain can be studied with NMR [49].

6.4.1.1 *In Vitro* Selection and Evolution Using Ribosome-Nascent Chain Complexes

RNCs provide a link between genotype (mRNA) and phenotype (protein) and thus can be used for *in vitro* peptide and protein selection experiments. Display techniques such as ribosome display [20] and mRNA-protein fusions [50] allow selecting for antibody single-chain Fv fragments (scFvs) and other proteins that interact with a molecule of interest. The starting library can encode up to 3×10^{11} different proteins which corresponds to a significantly larger library size compared to typical library sizes used for phage display selections (~10^7–10^8). Thus, the sequence space explored by *in vitro* selection is much larger compared to selection methods that involve a transformation or transfection step into a host cell. In ribosome display (Fig. 6.4), a DNA library is first transcribed using the T7 RNA polymerase and then translated *in vitro*. The mRNA sequences encoding the protein library do not contain a stop codon, but possess a long linker sequence which encodes for a C-terminal spacer peptide that spans the ribosomal exit tunnel. Therefore, the protein part is displayed outside of the ribosomal tunnel and can fold. The RNCs are then mixed with the protein of interest containing an affinity tag and the ribosomal complexes binding to it are therefore co-purified during the subsequent affinity purification. High Mg^{2+} concentration and low temperature allows preserving the ribosomal complexes such that the mRNA remains bound. After affinity purification, EDTA addition leads to disassembly of the RNCs and the release of the mRNAs. These mRNAs for selected binders and their sequences can be recovered and amplified by RT-PCR. The T7 promoter sequence is reintroduced during the PCR amplification step (Fig. 6.4).

The PCR product can then be used for further ribosome display cycles in order to enrich the best binders. Due to PCR errors the protein sequences can evolve *in vitro* during the selection

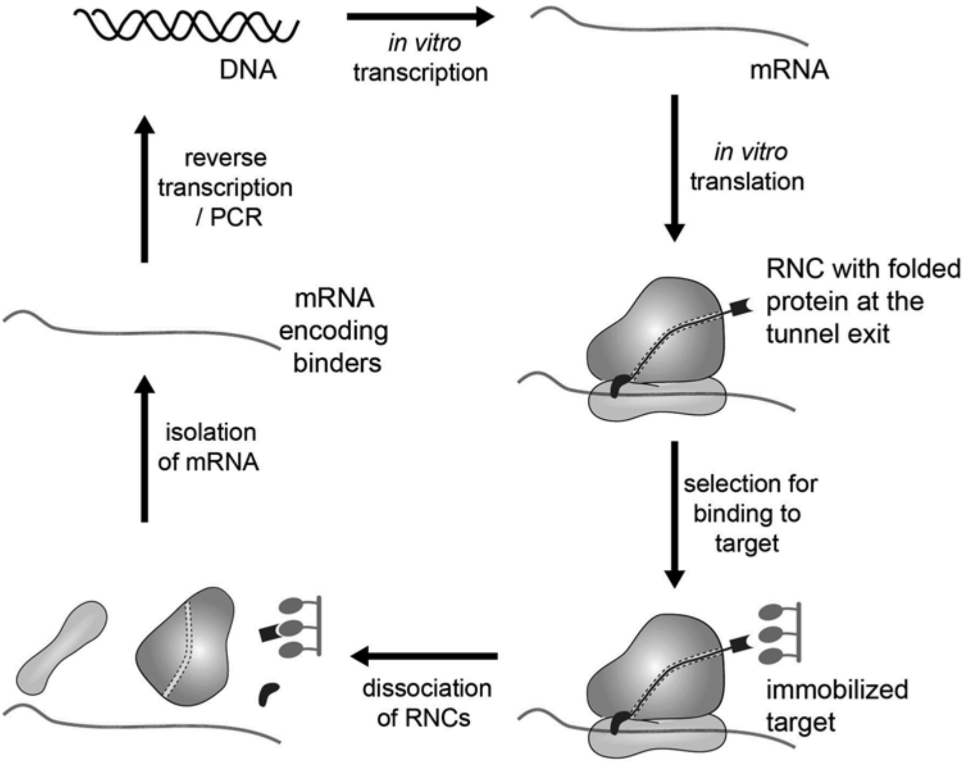


Fig. 6.4 ***In vitro*** **selection and evolution of protein by ribosome display**. A DNA library is transcribed and then translated *in vitro*. The mRNA sequences lack a stop codon and encode a linker sequence for a C-terminal peptide that spans the ribosomal exit tunnel. Therefore, the proteins encoded by the library can fold. Subsequently, RNCs are mixed with the immobilized target protein of interest. The RNCs interacting with the target protein are co-purified, while the others are washed away during the subsequent affinity purification. EDTA addition leads to dissociation of the RNCs and release of their mRNAs which can be recovered and amplified by RT-PCR. The T7 promoter sequence is reintroduced during the PCR amplification step. The resulting PCR products are subjected to further ribosome display cycles in order to enrich the best binders

experiment, and finally proteins with significantly improved affinity are selected which were not encoded by the original library pool [51]. Using PCR mutagenesis protocols, this can of course be exploited for *in vitro* evolution of proteins towards higher affinity, stability or in case of enzymes improved/altered substrate specificity. For these experiments, it is very important that at each step the diversity of the library is maintained: ideally, each member of the library is present in the experiment in several copies.

The concept of mRNA-protein fusions [50] is very similar to ribosome display. The major difference between the two methods is that a DNA spacer with a 3′ puromycin is fused to the mRNA encoding the protein library. During the *in vitro* translation reaction, the puromycin can enter the ribosome peptidyl transferase center, and subsequently the nascent polypeptide is transferred to puromycin. Thus, a covalent link is generated between the encoding mRNA and the protein allowing for harsher screening conditions compared to ribosome display where the intactness of the RNCs is crucial.

The ribosome display approach has been successfully used for the generation of high-affinity and highly specific scFvs [51, 52]. More recently, target proteins of bioactive small molecules (drugs) were selected by ribosome display from a library encoding full-length human proteins [53]. Ribosome display was very successfully applied to screen for Designed Ankyrin Repeat Proteins (DARPins), which are designed based on small, concave-shaped, α-helical protein domains typically involved in protein-protein interactions *in vivo*. The generation of DARPin libraries allows the selection of specific binders to virtually any protein of interest with up to low picomolar affinity. The stability of the core scaffold of DARPins leads to high-level expression and robust folding in ribosome display experiments. Indeed, issues exist with displaying scFvs because of their low folding efficiency. This is partially due to the disulfide bond that needs to be formed in the two immunoglobulin (Ig) domains. Cell-free transcription-translation is routinely performed under reducing conditions, while Ig domains require an oxidative environment for their folding. In ribosome display, this has been addressed by omitting reducing agents (DTT, Dithiothreitol) from the translation reaction and the addition of protein disulfide isomerase (PDI) for improved folding.

Antibody discovery and engineering is of high pharmaceutical interest. Accordingly, many groups developed cell-free expression-based tools to generate antibodies as diagnostics and drugs. For instance the use of the PURE system has several advantages [54] because of its low nuclease and protease activities as well as the absence of the tmRNA complex which increases the stability of the RNCs and allows screening of even larger libraries. The composition of the PURE reaction can be adjusted, release factors are omitted from the reaction, PDI and oxidized glutathione are added leading to proper folding of antibody fragments. A different construct design now also allows to screen libraries of Fabs (Fragment antigen-binding) which are usually more stable than scFvs [54].

6.4.2 Cell-Free Membrane Protein Expression

Membrane proteins represent about one third of the proteome of a cell. However, their study is often hampered by the lack of a suitable expression system. High-level overexpression of membrane proteins is frequently toxic for the cell. Moreover, the copy number of proteins is limited by the translocation and folding machinery as well as the space which is offered by the membrane bilayer of the host. Cell-free expression of membrane proteins allows overcoming several of these difficulties as it can be adapted to the expression of hydrophobic proteins.

Different possibilities exist to express membrane proteins in a cell-free expression system. First, it is possible to refold the precipitate which is formed during the cell-free expression of a membrane protein. This is achieved by solubilization of the aggregated proteins with detergent for a few hours under gentle agitation (precipitation-forming cell-free, P-CF) (Fig. 6.5). Not all detergents are suitable for the refolding step but

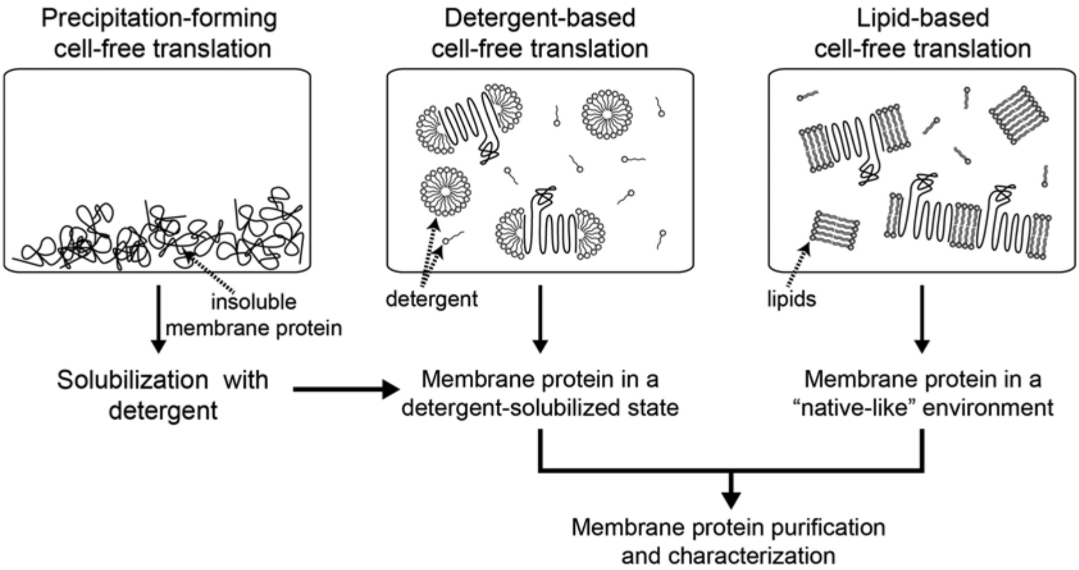

Precipitation-forming cell-free translation
Detergent-based cell-free translation
Lipid-based cell-free translation
insoluble membrane protein
detergent
lipids
Solubilization with detergent
Membrane protein in a detergent-solubilized state
Membrane protein in a "native-like" environment
Membrane protein purification and characterization

Fig. 6.5 Cell-free synthesis of membrane proteins. Three strategies are used to produce membrane proteins *in vitro*: in a conventional cell-free translation reaction the membrane protein precipitates (*left*). Subsequently, the aggregated protein is solubilized with detergent, in the presence of which it can fold into its correct structure. Several mild detergents can be added directly to the translation reaction without interfering with translation (*middle*), thus preventing the aggregation of the hydrophobic membrane proteins. In the presence of membranes, some membrane proteins can spontaneously insert into the lipid bilayer (*right*). The correct folding of the *in vitro* produced membrane proteins needs to be verified in functional assays

dodecylmaltoside (DDM), dodecylphosphocholine and lyso-phosphoglycerol derivatives (LMPG, LPPG) have been shown to successfully solubilize precipitates [55]. This approach has been successfully applied to the production of EmrE, a multidrug transporter [55, 56], and to the human histamine-1 receptor [57]. Second, the addition of detergent directly to the cell-free reaction keeps the nascent membrane proteins in solution (detergent-based cell-free, D-CF) (Fig. 6.5). Like in the P-CF approach, not all the detergents can be used in the detergent-based cell-free expression system. Detergents with a high critical micellar concentration (CMC) such as CHAPS have a tendency to destabilize the translation machinery. In contrast, mild detergents like DDM and digitonin are efficient for D-CF expression of EmrE [58]. Other surfactants which are traditionally not used for membrane solubilization because of their low efficiency to solubilize lipid bilayers have been shown to be particularly useful to stabilize membrane protein during the D-CF: MscL, the mechanosensitive channel, is efficiently expressed as a soluble protein in the presence of amphipols [59]. Compared to the cellular expression and the P-CF, the D-CF expression offers several advantages: (i) it avoids the formation of aggregates; (ii) it avoids the membrane integration step which is limited by the targeting and translocation efficiency, thus improving the production of the protein; (iii) the detergent-solubilized membrane protein can be used immediately. A third approach is based on the addition of lipids to the reaction mixture (lipid-based cell-free). Here, the classical cell-free reaction is supplemented with a preformed lipid bilayer (Fig. 6.5). This membrane-like environment can be either liposomes, bicelles or nanodiscs. While the membrane protein is synthesized at the ribosomes, the transmembrane segments are thought to spontaneously insert into the lipid bilayer offered by those lipidic environments. The main advantage of this technique is that the membrane protein will be produced in a "native-like" environment which is necessary to obtain a functional protein. Not only single membrane proteins can be prepared following these protocols. In fact, several membrane protein complexes have been generated using these methods. For instance, the F_1F_0-ATP synthase complex has been produced using the three techniques, P-CF, D-CF and L-CF [60]. Importantly, the complexes produced by the three cell-free expression protocols were similar to the *in vivo* complex in terms of enzymatic activity and structural properties. Using the L-CF approach, the SecYEG complex was produced *in vitro* [61]. Preformed liposomes were added to the reaction mixture and during translation, the SecYEG complex spontaneously inserted into the liposomes bilayer. The SecYEG translocon produced in this way was functionally active in the translocation of other membrane proteins.

Taken together, CFPS has been proven to be a very useful approach to overcome common problems faced with the traditional cellular expression system of membrane proteins.

6.4.3 Synthetic Biology

Synthetic biology is a rapidly expanding field which is currently actively researched. The idea to engineer biology in order to develop new biotechnological tools is indeed very attractive. Cell-free synthesis can be used to reproduce cellular pathways ex-vivo. On the one hand, the PURE system can allow deciphering the components required to realize a specific biological process. On the other hand, the classical cell-free extract can be the basis for the comprehensive synthesis and assembly of cellular macromolecules towards the development of a synthetic cell.

6.4.3.1 Bottom-Up Approach

Using the PURE system, it was possible to reconstitute bacterial transcription initiation from five different plasmids [62]. The α, β, β′ and ω subunits of the *E. coli* RNA polymerase as well as a σ factor (σ^{32} or σ^{70}) were co-expressed by the PURE machinery using T7 promoter. In this study, the expression and correct assembly of the RNA polymerase and the σ-factor-dependent transcription initiation was confirmed by production of luciferase from a linear DNA template under the control of an *E. coli* promoter [62]. It was found that the ω subunit is dispensable for transcription initiation. It is now possible to assess the activity of point mutants of the different subunits of the *E. coli* RNA polymerase. This work could not be performed in bacteria since the expressed variants are likely toxic to the cell. The work also paves the way to study the assembly and the function of other bacterial RNA polymerases for which we have little knowledge.

More recently, the co-expression of 13 genes building up a replication machinery was reported [63]. Step-by-step the authors produced a functional Pol III HE, which is composed of nine different proteins and forms an assembly of 17 subunits. Together with the primase DnaG, it was possible to replicate the G4 phage ssDNA. Using this remarkable system, it was demonstrated that all genes but *dnaQ*, a proofreading exonuclease, are required for replication activity. The initiation machinery consisting of DnaA possessing the initiator activity, DnaB helicase and DnaC, the helicase loader, was also produced in the PURE system [63]. It was demonstrated that these three proteins are essential and sufficient for initiation of replication. The authors were also able to reconstitute replication activity using a mixture of proteins/complexes produced in different tubes. It was possible to detect ssDNA replication using 13 genes (Pol III HE genes and *dnaA*, *dnaB*, *dnaC* and *dnaG*) when the PURE synthesis reaction was performed in a single tube. Moreover, the dsDNA produced by the neo-synthesized replication machinery possesses a biological activity as shown in a phage-plaque forming assay. Finally, a synthetic gene circuit using GFP as reporter showed the possibility to produce the complete and functional replication machinery producing a dsDNA containing GFP under the control of the T7 promoter, the only polymerase present in the PURE system. The final production of GFP confirmed the *in vitro* central dogma in a single tube [63].

6.4.3.2 Cell-Like Systems

A completely different strategy has been pursued for the development of a cell-free expression toolbox for synthetic biology [64]. A very simple approach based on bead-beater cell breaking was developed to prepare a reproducible, highly active S30 extract. High expression levels of eGFP were obtained under the control of the sigma factor 70, and therefore endogenous RNA polymerase was used for transcription [64]. The aim is to set up a close to native *E. coli* system to test synthetic gene circuits and to develop an artificial cell. This system enabled the assembly of the bacterial actin MreB on membranes after cell-free transcription/translation inside large liposomes [65]. Furthermore, it was shown that the presence of MreC is required to obtain filamentous structures (Fig. 6.6a, b). An organized cytoskeleton-like structure could thus be obtained inside liposome vesicles by using cell-free expression system producing MreB and MreC.

Large vesicles of more than 10 μm encapsulating the extract were formed using dispersion of small droplets in an oil phase as a first step [66]. Expression of α-hemolysin lasting for more than 4 days was achieved in this system by exchange of small, up to 3 kDa molecules across the membrane bilayer leading to a continuous supply of substrates for the transcription and translation reactions. This system is therefore the first step towards a bioreactor encapsulated inside a lipid vesicle and able to express proteins for more than 4 days. A step forward was achieved by the expression of the whole T7 bacteriophage genome, containing about 60 genes encoded by 40 kbp DNA. The complete proteome was synthesized using an *E. coli* cell-free transcription-translation system. Billions of T7 bacteriophages, assembled spontaneously into well-shaped particles (Fig. 6.6c, d), are produced per milliliter of batch reaction. Importantly, these *in vitro* assembled phages are as infectious as *in vivo* synthesized ones [67].

This approach opens up the possibility to directly and rapidly assess genetic circuits and the effects of promoter strength or different substrate concentrations, to help understanding bacterial cell metabolism. Very recently, two-dimensional DNA compartments in silicon were generated [68]. In these compartments protein expression cycles can be auto-regulated using interconnected compartments containing different sets of DNA. This approach aims to study biological networks and communication between cells.

6.4.3.3 Expansion of the Genetic Code

A clear advantage of the cell-free expression is the possibility to efficiently and specifically synthesize proteins with non-natural amino acids. It is possible to replace a certain amino acid by a non-natural analogue provided that the corresponding amino acyl t-RNA synthase (aaRS) recognizes the unnatural amino acid. This can be easily achieved for seleno-methionine which is used in crystallography to solve the phase problem [69] and to structurally similar analogues of proline, tyrosine, phenylalanine, leucine and valine (reviewed in [70]). To further expand the repertoire of amino acids, stop codon suppressor-tRNAs were employed that recognize the amber stop codon were chemically acylated with artificial amino acids [71]. The advantage of this approach is that the incorporation of the artificial amino acid is site specific. Similarly, pairs of specific tRNAs—recognizing the amber stop codon or even a 4-base codon—and engineered aaRSs were evolved to incorporate the artificial amino acid at a specific site of the protein. Several tRNA/aaRS pairs are required to incorporate two or more unnatural amino acids in one protein for protein folding studies using FRET (*e.g.*, [72]). This represents a very powerful approach for investigation of protein structure, function and dynamics. To improve the efficiency of stop codon suppression, release factor RF1 can be omitted from the cell-free translation reaction. For improved 4-base codon tRNA recognition, 'orthogonal' ribosomes have been engineered [73]. Similarly, an engineered elongation factor EF-Tu exhibiting improved affinity for incorporation of phosphoserine was reported.

The application possibilities of such unnatural proteins are manifold: ranging from protein folding and protein-protein interaction studies using amino acids with fluorescent dyes or photo-activatable crosslinkers, to production of protein conjugates with small molecules or synthetic polymers for protein therapeutics. Of particular

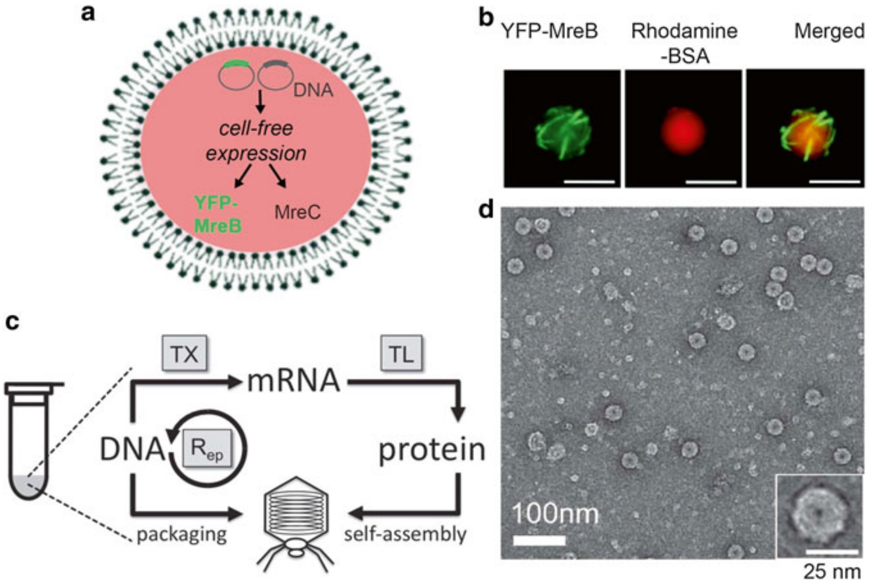


Fig. 6.6 Successful examples of cell-free synthetic biology. (**a**) Scheme of cell-free co-expression of YFP-tagged MreB and MreC inside a liposome. (**b**) Expression of the YFP-MreB fusion protein together with MreC results in the formation of filamentous structures (*left panel*), rhodamine-BSA stains the lumen of the lipid vesicle (*middle panel*). The merged red and green image highlights the localization of the YFP-MreB filament on the surface of the liposome. The scale bar corresponds to 10 μm. (**c**) General scheme of the coupled *in vitro* transcription-translation reaction allowing the production of assembled and infectious phage particles from the complete 40 kbp genome. (**d**) Transmission electron microscope micrograph of PHIX174 phage particles produced by the cell-free system. *Inset*: close-up view of an *in vitro* synthesized phage (Panels **a** and **b** are adapted with permission from Ref. [65]. Panels **c** and **d** are adapted with permission from Ref. [67])

interest are antibody–drug conjugates and polyethylene glycol-growth factor conjugates with improved bio-kinetics [74].

6.5 Limitations

In the case of large scale expression for structural studies, the main limitation of *E. coli* cell free extract resides in the cost of the chemicals that have to be added to the system. Furthermore, the use of bacterial extracts leads to the production of proteins without any post-translational modifications, which are sometimes crucial for proper folding and function of eukaryotic proteins. *E. coli* cell-free expression is therefore most successful for expression of bacterial proteins. Eukaryotic cell-free expressions are often rather inefficient, resulting in low protein yields—this is most likely due to the lack of translation factors in the cell extracts. Moreover, eukaryotic cell-free expression systems are more labor-intensive, for instance requiring capped and polyadenylated mRNA for *in vitro* efficient translation.

For expression of protein complexes, cell-free expression is limited to bacterial or phage protein complex expression, notably because of the limited protein size that can be expressed in *E. coli* per se (proteins larger than 100 kDa are difficult to produce in *E. coli*). This also applies when the transcription-translation machinery is “purified”. In general, the ribosomal machinery tends to be less efficient as soon as it is extracted from the cell and even more in the case of the PURE system in which it has been shown that the ribosomes are ten-times slower than the ones in the cell, incorporating only two amino acids per second [62]. Furthermore, the different enzymes including the ribosomes become less active over time outside of the cell. Due to this limited efficiency, cell-free protein expression did not become a general method for protein production.

6.6 Outlook

As highlighted in this review, cell-free expression is particularly suited for specific structural biology applications, *in vitro* protein screening, selection and evolution as well as for synthetic biology. One main advantage is the possibility of specific protein labeling, for instance, in NMR and the possibility to incorporate unnatural amino acids at specific sites of the protein. Here, we provide several examples that apply cell-free expression to produce large assemblies including phages. In these cases, the cell-free systems are used for production of small quantities for analytical purposes and functional studies, rather than large scale protein production.

Cell-free translation is routinely used to study the translation process itself. Recent advances in single molecule techniques may even allow following co-translational processes such as protein folding during active protein synthesis, rather than using stalled RNCs.

For structural biology, cell-free production of complexes comes to the fore when ribosomal complexes are studied. To date, cell-free extracts from eukaryotic species such as yeast, wheat germ, insect cells, rabbit reticulocytes and HeLa cells are constantly improved for protein production. A reconstituted system has been reported for the study of the mechanisms of mammalian protein synthesis [75]. With these cell-free systems, specific eukaryotic RNC complexes can be generated *in vitro* and structurally and functionally characterized to understand the complex function of the eukaryotic translation machinery.

Acknowledgements C.S. acknowledges funding by the European Research Council Starting grant (ComplexNMD, 281331); and by a Sinergia grant from the Swiss National Science Foundation (CRSII3_136254/1).

References

1. Borsook H (1950) Protein turnover and incorporation of labeled amino acids into tissue proteins *in vivo* and *in vitro*. Physiol Rev 30(2):206–219
2. Winnick T (1950) Studies on the mechanism of protein synthesis in embryonic and tumor tissues. I. Evidence relating to the incorporation of labeled amino acids into protein structure in homogenates. Arch Biochem 27(1):65–74
3. Palade GE (1955) A small particulate component of the cytoplasm. J Biophys Biochem Cytol 1(1):59–68
4. Tissieres A, Schlessinger D, Gros F (1960) Amino acid incorporation into proteins by *Escherichia coli* ribosomes. Proc Natl Acad Sci U S A 46(11):1450–1463
5. Nirenberg MW (1963) Cell-free protein synthesis directed by messenger RNA. Methods Enzymol 6:17–23
6. Nirenberg MW, Matthaei JH (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A 47:1588–1602
7. Zubay G (1973) *In vitro* synthesis of protein in microbial systems. Annu Rev Genet 7:267–287
8. Pratt JM (1984) Coupled transcription–translation in prokaryotic cellfree systems. In: Hemes BD, Higgins SJ (eds) Current protocols. IRL Press, Oxford, pp 179–209
9. Lesley SA (1995) Preparation and use of *E. coli* S-30 extracts. Methods Mol Biol 37:265–278
10. Spirin AS (1991) Cell-free protein synthesis bioreactor. Front Bioprocess II:31–43
11. Spirin AS, Baranov VI, Ryabova LA, Ovodov SY, Alakhov YB (1988) A continuous cell-free translation system capable of producing polypeptides in high yield. Science 242(4882):1162–1164
12. Kigawa T et al (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett 442(1):15–19
13. Kigawa TIM, Aoki M, Matsuda T, Yabuki T, Seki E, Harada T, Watanabe S, Yokoyama S (2008) The use of the *Escherichia coli* cell-free protein synthesis for structural biology and structural proteomics. In: Spirin AS, Swartz JR (eds) Cell-free protein synthesis: methods and protocols. Wiley, New York, pp 99–109
14. Vinarov DA et al (2004) Cell-free protein production and labeling protocol for NMR-based structural proteomics. Nat Methods 1(2):149–153
15. Yokoyama S et al (2000) Structural genomics projects in Japan. Nat Struct Biol 7(Suppl):943–945
16. He F et al (2009) Structural and functional characterization of the NHR1 domain of the Drosophila neuralized E3 ligase in the notch signaling pathway. J Mol Biol 393(2):478–495
17. Koglin A et al (2006) Combination of cell-free expression and NMR spectroscopy as a new approach for

structural investigation of membrane proteins. Magn Reson Chem 44 Spec No:S17–S23

18. Maslennikov I et al (2010) Membrane domain structures of three classes of histidine kinase receptors by cell-free expression and rapid NMR analysis. Proc Natl Acad Sci U S A 107(24):10902–10907
19. Yokoyama S (2003) Protein expression systems for structural genomics and proteomics. Curr Opin Chem Biol 7(1):39–43
20. Hanes J, Plückthun A (1997) *In vitro* selection and evolution of functional proteins by using ribosome display. Proc Natl Acad Sci U S A 94(10):4937–4942
21. Klammt C et al (2007) Cell-free production of G protein-coupled receptors for functional and structural studies. J Struct Biol 158(3):482–493
22. Chen YJ et al (2007) X-ray structure of EmrE supports dual topology model. Proc Natl Acad Sci U S A 104(48):18999–19004
23. Deniaud A, Liguori L, Blesneac I, Lenormand JL, Pebay-Peyroula E (2010) Crystallization of the membrane protein hVDAC1 produced in cell-free system. Biochim Biophys Acta 1798(8):1540–1546
24. Boland C et al (2014) Cell-free expression and in meso crystallisation of an integral membrane kinase for structure determination. Cell Mol Life Sci 71(24):4895–4910
25. Kim TW et al (2006) Simple procedures for the construction of a robust and cost-effective cell-free protein synthesis system. J Biotechnol 126(4):554–561
26. Yang HL, Zubay G (1973) Synthesis of the arabinose operon regulator protein in a cell-free system. Mol Gen Genet 122(2):131–136
27. Ryabova LA, Desplancq D, Spirin AS, Plückthun A (1997) Functional antibody production using cell-free translation: effects of protein disulfide isomerase and chaperones. Nat Biotechnol 15(1):79–84
28. Iskakova MB, Szaflarski W, Dreyfus M, Remme J, Nierhaus KH (2006) Troubleshooting coupled *in vitro* transcription-translation system derived from *Escherichia coli* cells: synthesis of high-yield fully active proteins. Nucleic Acids Res 34(19), e135
29. Shimizu Y et al (2001) Cell-free translation reconstituted with purified components. Nat Biotechnol 19(8):751–755
30. Fedorov AN, Baldwin TO (1998) Protein folding and assembly in a cell-free expression system. Methods Enzymol 290:1–17
31. Von Loeffelholz O, Botte M, Schaffitzel C (2011) *Escherichia coli* cotranslational targeting and translocation. ELS. doi:10.1002/9780470015902.a0023170
32. Walter P, Blobel G (1983) Preparation of microsomal membranes for cotranslational protein translocation. Methods Enzymol 96:84–93
33. Müller M, Blobel G (1984) *In vitro* translocation of bacterial proteins across the plasma membrane of *Escherichia coli*. Proc Natl Acad Sci U S A 81(23):7421–7425
34. Schulze RJ et al (2014) Membrane protein insertion and proton-motive-force-dependent secretion through the bacterial holo-translocon SecYEG-SecDF-YajC-YidC. Proc Natl Acad Sci U S A 111(13):4844–4849
35. Yi L, Celebi N, Chen M, Dalbey RE (2004) Sec/SRP requirements and energetics of membrane insertion of subunits a, b, and c of the *Escherichia coli* F1F0 ATP synthase. J Biol Chem 279(38):39260–39267
36. Schaffitzel C, Ban N (2007) Generation of ribosome nascent chain complexes for structural and functional studies. J Struct Biol 158(3):463–471
37. Giudice E, Mace K, Gillet R (2014) Trans-translation exposed: understanding the structures and functions of tmRNA-SmpB. Front Microbiol 5:113
38. Bischoff L, Berninghausen O, Beckmann R (2014) Molecular basis for the ribosome functioning as an L-tryptophan sensor. Cell Rep 9(2):469–475
39. Merz F et al (2008) Molecular mechanism and structure of Trigger Factor bound to the translating ribosome. EMBO J 27(11):1622–1632
40. Schaffitzel C et al (2006) Structure of the E. coli signal recognition particle bound to a translating ribosome. Nature 444(7118):503–506
41. Estrozi LF, Boehringer D, Shan SO, Ban N, Schaffitzel C (2011) Cryo-EM structure of the E. coli translating ribosome in complex with SRP and its receptor. Nat Struct Mol Biol 18(1):88–90
42. von Loeffelholz O et al (2013) Structural basis of signal sequence surveillance and selection by the SRP-FtsY complex. Nat Struct Mol Biol 20(5):604–610
43. Zhang X, Schaffitzel C, Ban N, Shan SO (2009) Multiple conformational switches in a GTPase complex control co-translational protein targeting. Proc Natl Acad Sci U S A 106(6):1754–1759
44. Frauenfeld J et al (2011) Cryo-EM structure of the ribosome-SecYE complex in the membrane environment. Nat Struct Mol Biol 18(5):614–621
45. Kohler R et al (2009) YidC and Oxa1 form dimeric insertion pores on the translating ribosome. Mol Cell 34(3):344–353
46. Clark PL, Ugrinov KG (2009) Measuring cotranslational folding of nascent polypeptide chains on ribosomes. Methods Enzymol 466:567–590
47. Hsu ST et al (2007) Structure and dynamics of a ribosome-bound nascent chain by NMR spectroscopy. Proc Natl Acad Sci U S A 104(42):16516–16521
48. Hsu ST, Cabrita LD, Fucini P, Christodoulou J, Dobson CM (2009) Probing side-chain dynamics of a ribosome-bound nascent chain using methyl NMR spectroscopy. J Am Chem Soc 131(24):8366–8367
49. O'Brien EP, Christodoulou J, Vendruscolo M, Dobson CM (2012) Trigger factor slows co-translational folding through kinetic trapping while sterically protecting the nascent chain from aberrant cytosolic interactions. J Am Chem Soc 134(26):10920–10932
50. Roberts RW, Szostak JW (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. Proc Natl Acad Sci U S A 94(23):12297–12302

51. Hanes J, Schaffitzel C, Knappik A, Pluckthun A (2000) Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. Nat Biotechnol 18(12):1287–1292
52. Schaffitzel C et al (2001) *In vitro* generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. Proc Natl Acad Sci U S A 98(15):8572–8577
53. Wada A, Hara S, Osada H (2014) Ribosome display and photo-cross-linking techniques for *in vitro* identification of target proteins of bioactive small molecules. Anal Chem 86(14):6768–6773
54. Kanamori T, Fujino Y, Ueda T (2014) PURE ribosome display and its application in antibody technology. Biochim Biophys Acta 1844(11):1925–1932
55. Klammt C et al (2004) High level cell-free expression and specific labeling of integral membrane proteins. Eur J Biochem 271(3):568–580
56. Elbaz Y, Steiner-Mordoch S, Danieli T, Schuldiner S (2004) *In vitro* synthesis of fully functional EmrE, a multidrug transporter, and study of its oligomeric state. Proc Natl Acad Sci U S A 101(6):1519–1524
57. Sansuk K et al (2008) GPCR proteomics: mass spectrometric and functional analysis of histamine H1 receptor after baculovirus-driven and *in vitro* cell free expression. J Proteome Res 7(2):621–629
58. Klammt C et al (2005) Evaluation of detergents for the soluble expression of alpha-helical and beta-barrel-type integral membrane proteins by a preparative scale individual cell-free expression system. FEBS J 272(23):6024–6038
59. Park KH et al (2007) Fluorinated and hemifluorinated surfactants as alternatives to detergents for membrane protein cell-free synthesis. Biochem J 403(1):183–187
60. Matthies D et al (2011) Cell-free expression and assembly of ATP synthase. J Mol Biol 413(3):593–603
61. Matsubayashi H, Kuruma Y, Ueda T (2014) *In vitro* synthesis of the E. coli Sec translocon from DNA. Angew Chem Int Ed Engl 53(29):7535–7538
62. Asahara H, Chong S (2010) *In vitro* genetic reconstruction of bacterial transcription initiation by coupled synthesis and detection of RNA polymerase holoenzyme. Nucleic Acids Res 38(13), e141
63. Fujiwara K, Katayama T, Nomura SM (2013) Cooperative working of bacterial chromosome replication proteins generated by a reconstituted protein expression system. Nucleic Acids Res 41(14):7176–7183
64. Shin J, Noireaux V (2010) Efficient cell-free expression with the endogenous E. Coli RNA polymerase and sigma factor 70. J Biol Eng 4:8
65. Maeda YT et al (2012) Assembly of MreB filaments on liposome membranes: a synthetic biology approach. ACS Synth Biol 1(2):53–59
66. Noireaux V, Libchaber A (2004) A vesicle bioreactor as a step toward an artificial cell assembly. Proc Natl Acad Sci U S A 101(51):17669–17674
67. Shin J, Jardine P, Noireaux V (2012) Genome replication, synthesis, and assembly of the bacteriophage T7 in a single cell-free reaction. ACS Synth Biol 1(9):408–413
68. Karzbrun E, Tayar AM, Noireaux V, Bar-Ziv RH (2014) Synthetic biology. Programmable on-chip DNA compartments as artificial cells. Science 345(6198):829–832
69. Watanabe M et al (2010) Cell-free protein synthesis for structure determination by X-ray crystallography. Methods Mol Biol 607:149–160
70. Hendrickson TL, de Crecy-Lagard V, Schimmel P (2004) Incorporation of nonnatural amino acids into proteins. Annu Rev Biochem 73:147–176
71. Noren CJ, Anthony-Cahill SJ, Griffith MC, Schultz PG (1989) A general method for site-specific incorporation of unnatural amino acids into proteins. Science 244(4901):182–188
72. Wang K et al (2014) Optimized orthogonal translation of unnatural amino acids enables spontaneous protein double-labelling and FRET. Nat Chem 6(5):393–403
73. Chin JW (2014) Expanding and reprogramming the genetic code of cells and animals. Annu Rev Biochem 83:379–408
74. Kim CH, Axup JY, Schultz PG (2013) Protein conjugation with genetically encoded unnatural amino acids. Curr Opin Chem Biol 17(3):412–419
75. Alkalaeva EZ, Pisarev AV, Frolova LY, Kisselev LL, Pestova TV (2006) *In vitro* reconstitution of eukaryotic translation reveals cooperativity between release factors eRF1 and eRF3. Cell 125(6):1125–1136

7 A *Bacillus megaterium* System for the Production of Recombinant Proteins and Protein Complexes

Rebekka Biedendieck

Abstract

For many years the Gram-positive bacterium *Bacillus megaterium* has been used for the production and secretion of recombinant proteins. For this purpose it was systematically optimized. Plasmids with different inducible promoter systems, with different compatible origins, with small tags for protein purification and with various specific signals for protein secretion were combined with genetically improved host strains. Finally, the development of appropriate cultivation conditions for the production strains established this organism as a bacterial cell factory even for large proteins. Along with the overproduction of individual proteins the organism is now also used for the simultaneous coproduction of up to 14 recombinant proteins, multiple subsequently interacting or forming protein complexes. Some of these recombinant strains are successfully used for bioconversion or the biosynthesis of valuable components including vitamins. The titers in the g per liter scale for the intra- and extracellular recombinant protein production prove the high potential of *B. megaterium* for industrial applications. It is currently further enhanced for the production of recombinant proteins and multi-subunit protein complexes using directed genetic engineering approaches based on transcriptome, proteome, metabolome and fluxome data.

Keywords

Bacillus megaterium • Recombinant proteins • Coexpression • Protein interaction • Multi-subunit protein complexes • Omic-technologies

R. Biedendieck (✉)
Institute of Microbiology, Technische Universität
Braunschweig, Braunschweig, Germany

Braunschweig Integrated Centre of Systems Biology
(BRICS), Technische Universität Braunschweig,
Braunschweig, Germany
e-mail: r.biedendieck@tu-braunschweig.de

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_7

7.1 Introduction

7.1.1 *Bacillus megaterium*

The Gram-positive bacterium *Bacillus megaterium* is a rod-shaped bacterium which was first described by de Bary more than 130 years ago [1]. With a volume 100 times as big as that of *Escherichia coli*, *B. megaterium* belongs to the larger eubacteria [2]. So, it got its name from the greek word "megatherium" meaning "big beast" (Fig. 7.1). The extraordinary dimensions of its vegetative cells and also of its spores made *B. megaterium* already in the 1960s a model organism for analyses of cell structure, of cell wall and cytoplasmic membrane synthesis as well as of spore formation [3, 4].

B. megaterium is mainly found in the soil, its major natural habitat, but the bacterium was also found in other, sometimes extreme environments including sweet honey or dried meat [4]. Other strains were isolated from seawater [5], sediments and even from fish [4]. Due to its broad environmental distribution, the organism is able to metabolize a large variety of carbon sources and possesses a high osmotic tolerance. These abilities allows this organism to grow on waste materials and low cost substances as raw glycerol, bagasse or molasses [6–8]. The genome sequences of at least three different *B. megaterium* strains are available [9, 10]. The genome size with around 5.100 Mbp, the genomic G+C content of 38.2 %, the 16S rRNA phylogeny as well as the phenotypic behavior place *B. megaterium* in phylogenetic trees comparable related to *Bacillus subtilis* and *Bacillus cereus* [9].

Although *B. megaterium* currently does not keep the GRAS (Generally Recognized As Safe) status like *B. subtilis* and *Bacillus licheniformis* do, it is completely nonpathogenic and consequently classified to the security level 1 (2013, Federal Ministry of Food and Agriculture, Germany). In contrast to Gram-negative organisms like *E. coli*, the Gram-positive *B. megaterium* lacks outer membrane associated endotoxins. Further on there are strains available which fail to form heat, organic solvents and lytic enzyme resistant spores. Other constructed security strains are sensitive against UV light or are strongly reduced in their DNA-repair mechanism [11–13]. These features make *B. megaterium* very well applicable in food and even in pharmaceutical industry [2, 4].

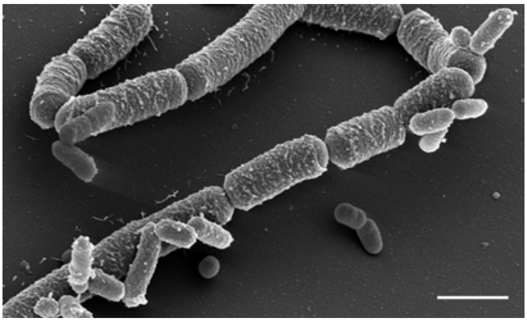


Fig. 7.1 Electron microscope image of vegetative cells of *Bacillus megaterium* (large cells) and *Escherichia coli* (small cells). *B. megaterium* and *E. coli* were individually grown in rich medium at 37 °C with strong shaking until reaching the middle of their exponential growth phases. Both cultures were mixed in a ratio of 1:1, cells were fixed with aldehyde and dehydrated with a graded series of acetone. After critical-point-drying with liquid CO_2 they were sputter-coated with gold and examined in a field emission scanning electron microscope (FESEM) Zeiss DSM982 Gemini at an acceleration voltage of 5 kV (SE in-lens detector and Everhart-Thronley SE-detector, 50:50 ratio). *B. megaterium* is reaching an up to 100 times larger volume compared to that of *E. coli* [2] (The magnification was 6,500-fold. The *white bar* corresponds to 2 μm. The picture was taken by Manfred Rohde, Helmholtz Centre for Infection Research, Braunschweig, Germany)

7.1.2 Plasmids of *Bacillus megaterium*

Plasmids, extrachromosomal DNA molecules that replicate independently from the genomic DNA, are found wide spread throughout bacteria [14]. With length spanning from only a few hundred to several hundred thousand base pairs they strongly vary in their dimensions and their gene content. Additional to that strong variations in their copy number are found [15]. Typically, the genetic machinery of the corresponding host is used at least in part for plasmid replication. The plasmid encoded genes often provide the host with additional capabilities, *i.e.* tolerance to otherwise toxic compounds or the metabolization of additional carbon or nitrogen sources.

Several *B. megaterium* strains are also known to carry significant parts of their genetic material on plasmids [4, 9]. A well characterized and genome sequenced *B. megaterium* strain harboring several plasmids is QM B1551. Its seven endogenous plasmids show copy numbers between 1 and 18 [9] and comprise approximately 11 % of the total cellular DNA [16]. The size spectrum of those plasmids reaches from only 5.4 to over 164 kb [4, 9]. The two smallest of these plasmids replicate by the rolling circle mechanism, whereas the other five plasmids are using the theta replication mechanism with cross-hybridization replicons [9, 17]. The DNA sequence of the rolling circle replicons show similarities to DNA sequences of plasmids known from *Bacillus thuringiensis* and *Bacillus anthracis* [18], while the five theta replicons appear to be unique, forming a new class of compatible replicons [19]. The plasmids of *B. megaterium* QM B1551 carry several interesting genes, such as genes coding for proteins involved in cell division, in germination, in heavy metal resistance, in cell wall hydrolysis or in rifampin resistance. Even a complete rRNA operon and 18 additional tRNA genes are located on one of these plasmids [18, 19]. *B. megaterium* strain PV361, a derivative of strain QM B1551 [20], lacks all seven plasmids. Surprisingly, this plasmidless derivative is able to grow similarly to its parental strain on rich medium and shows no differences in sporulation. Therefore, the plasmids of QM B1551 may play a role in the adaptation to various special environmental conditions including the presence of heavy metals or antibiotics [2, 4].

For industrial applications and research, plasmidless strains are used as hosts for the efficient plasmid-driven production of recombinant proteins. Unlike strain QM B1551, *B. megaterium* strain DSM319 naturally does not carry any plasmid making it a well suited host for the production of plasmid-encoded recombinant proteins [9]. Furthermore, in contrast to *B. subtilis*, *B. megaterium* is known to stably replicate and maintain recombinant plasmids for a long time even without the selective pressure of antibiotics [4, 21–23]. Moreover, an efficient protocol for protoplast transformation as well as protocols for transconjugation, gene knock-out and replacements are established [24].

Most of the plasmids used for the production of recombinant proteins in *B. megaterium* are based on the replicon of the *B. cereus* plasmid pBC16 *oriU* / *repU* (Table 7.1) [21, 22, 25]. Additionally, the replicons of two (*repM100*, *repM700*) of the seven plasmids of *B. megaterium* strain QM B1551 were used to construct further vectors utilizable for the recombinant protein production in *B. megaterium* [9, 16, 17, 26–28]. As all replicons of the QM B1551 plasmids are compatible with that of pBC16, the recombinant plasmids based on *repM100* and *repM700*, respectively, can be used for coexistence with pBC16-derivatives allowing coproduction of recombinant proteins in *B. megaterium* [26, 27].

7.2 Recombinant Protein Production in *Bacillus megaterium*

For many decades *E. coli* was systematically developed to the currently best studied and mostly used bacterial host for recombinant gene expression [29]. However, its Gram-negative character can sometimes limit the use as a production host for recombinant proteins. Its outer membrane mostly impedes an efficient release of target proteins into the culture supernatant rather detaining them either intracellularly or in the periplasm. Some recombinant proteins tend to form insoluble protein aggregates in the cytoplasm or are simply toxic to *E. coli* [29]. Moreover, *E. coli* reveals some problems during the production of proteins with larger molecular weight above 200,000 Da. These observation argue for the need to develop alternative expression systems like *B. megaterium* turns out to be [2, 4, 29–31].

Table 7.1 Comparison of intracellularly produced green fluorescent protein (GFP) from recombinant *B. megaterium* with plasmids carrying different promoters and different origins of replication/replicons at different growth conditions

Promoter	Origin of replication/replicon	Growth condition	Maximal GFP amount	References
Native P_{xylA}	*oriU*	Shake flasks	17.9 mg/L / 14 mg/g_{CDW}	[47]
Native P_{xylA}	*oriU*	HCDC	274 mg/L / 5.2 mg/g_{CDW}	[47]
Native P_{xylA}	*repM100*	Shake flasks	13.9 mg/L / 7.8 mg/g_{CDW}	[26]
Native P_{xylA}	*repM700*	Shake flasks	3.5 mg/L / n. d.	[26]
2× native P_{xylA} (two vectors)	*repM100 / oriU*	Shake flasks	16.4 mg/L / 14.6 mg/g_{CDW}	[26]
Optimized P_{xylA}	*oriU*	Shake flasks	124 mg/L / 82.5 mg/g_{CDW}	[40]
Optimized P_{xylA}	*oriU*	Fed-batch	1.25 g/L / 36.8 mg/g_{CDW}	[40]
P_{sacB}	*oriU*	Shake flasks	5.5 mg/L / 7.9 mg/g_{CDW}	[46]
P_{T7} (two vectors)	*oriU/repM100*	Shake flasks	50 mg/L / 42.3 mg/g_{CDW}	[27]
P_{K1E} (two vectors)	*oriU/repM100*	Shake flasks	n.d. / 38.1 mg/g_{CDW}	[44]
P_{SP6} (two vectors)	*oriU/repM100*	Shake flasks	n.d. / 4.3 mg/g_{CDW}	[44]

The amount of GFP was measured spectroscopically and is presented as volumetric (mg/L) and cellular (mg/g_{CDW}) amount

oriU Origin of replication from the plasmid pBC16 from *B. cereus* [25], *repM100* replicon of plasmid pBM100 from *B. megaterium* strain QM B1551 [9], *repM700* replicon of plasmid pBM700 from *B. megaterium* strain QM B1551 [26], P_{xylA} xylose-inducible promoter from *B. megaterium*, P_{T7} T7-RNA-polymerase dependent promoter from the phage T7, P_{sacB} sucrose-inducible promoter from *B. megaterium*, P_{K1E} K1E-RNA-polymerase dependent promoter from the *E. coli* phage K1E, P_{SP6} SP6-RNA-polymerase dependent promoter from the *Salmonella typhimurium* phage SP6, *HCDC* high cell density cultivation, *CDW* cell dry weight, *n.d.* not determined

7.2.1 Promoter Systems for the Recombinant Protein Production in *Bacillus megaterium*

The beginnings of *B. megaterium* as a host for plasmid-based recombinant protein production were summarized in great detail [3]. The expression of the recombinant genes was mostly driven by the native promoters of the corresponding genes. However, the controlled high level production of recombinant proteins and multisubunit protein complexes requires strong inducible promoters to regulate the process and to reduce the metabolic burden to the bacterial cell [32]. Here, the scientific challenge lies in the control of the high basic activity of many strong promoters and in the discovery of new inducible promoters as well as in the appropriate combination of them.

7.2.1.1 The Xylose-Inducible Promoter System

Certainly, the development of the xylose-inducible promoter system for the use of *B. megaterium* in recombinant protein production was the first groundbreaking step. In the genome of *B. megaterium*, the operon for xylose uptake and utilization is located divergently transcribed to the gene encoding its repressor XylR. Interestingly, the promoter regions of the operon (P_{xylA}) and that of the regulator gene (P_{xylR}) have overlapping regulatory elements [21, 33]. Currently, the theory is that in the absence of xylose the repressor XylR is bound to the two tandem overlapping operators of P_{xylA} preventing transcription of the operon [34, 35]. In the presence of xylose the sugar binds to the repressor protein resulting in a conformational change. This new conformation of XylR does not efficiently bind the operator regions. The steric hampering of RNA polymerase binding is eliminated and the transcription of the operon can occur. An additional level of regulation is mediated by so called catabolite repression [36]. A catabolite control sequence (*cre*) is located in the reading frame of the first gene in the operon. In the presence of glucose the catabolite control protein CcpA binds to this *cre*-element also reducing the expression of the operon even when xylose is present. Hence, in the presence of xylose and glucose operon expression is decreased 14-fold [36].

Besides glucose, also fructose, mannitol, arabinose, glycerol and ribose were found to cause repression but to a lesser extent compared to glucose [36].

For the construction of a plasmid-based xylose-inducible gene expression system, P_{xylA} with the first 195 bp of the following gene *xylA* as well as P_{xylR} with the gene encoding the repressor XylR were cloned into a *Bacillus*-*E. coli*-shuttle vector based on the *B. cereus* plasmid pBC16 and the *E. coli* plasmid pBR322 [21, 25, 37]. The reduced promoter activity caused by catabolite repression was overcome by eliminating the *cre*-DNA sequence. Instead a multiple cloning site (MCS) consisting of 15 unique restriction sites was introduced downstream of the first 15 bp of *xylA* [38]. These adapted DNA sequences of promoter and MCS were also introduced into plasmids carrying the replicons *repM100* and *repM700* suitable for coexistence with the replicon of pBC16 in *B. megaterium* [27]. Meanwhile also vectors are available for *B. megaterium*, which allow restriction enzyme free cloning (gateway cloning) [39]. Further on, by individually or in parallel changing the sequences of the −10 and −35 box of P_{xylA}, the untranslated DNA-region upstream of the *xylA* start codon including the XylR-binding site as well as the ribosome binding site (*rbs*) to predicted consensus sequences, the expression efficiency of the original xylose-inducible system was improved up to 18-fold [40]. Further on, the promoter strength can be regulated from zero to full function by addition of different amounts of xylose [21].

7.2.1.2 Phage RNA-Polymerase Driven Promoters

Typically, prokaryotic promoters are recognized by the host RNA-polymerase which then transcribes the following gene or operon into the corresponding mRNA. This is also true for recombinant promoters located in the genome or on plasmids. In 1985, Tabor and Richardson described the first bacterial expression system which was based on the RNA polymerase of the bacteriophage T7 (T7 RNAP) in combination with its corresponding promoter P_{T7} and terminator [41]. This system is characterized by a high processivity and stringent selectivity towards its native promoter making it very suitable for recombinant expression. After its first discovery and use in *E. coli* [41] the system was adopted to further Gram-negative and also Gram-positive bacteria [42, 43].

In 2009 and 2010, different bacteriophage driven systems were also developed for *B. megaterium*. The corresponding pairs of promoter and terminator were individually introduced into *repU* dependent *Bacillus*-*E. coli*-shuttle vectors around an enhanced MCS also found in other *B. megaterium* vectors [27, 44]. The genes for RNA polymerase from the well-known bacteriophage T7 (T7 RNAP), the *E. coli* phage K1E (K1E RNAP) and the *Salmonella typhimurium* phage SP6 (SP6 RNAP) were individually cloned into *Bacillus*-*E. coli*-shuttle vectors carrying the *repM100* replicon compatible with the plasmids carrying the corresponding promoters. The expression of the phage RNA polymerase encoding genes were set under control of the native xylose-inducible promoter [27, 44]. When comparing the strength of the native bacterial P_{xylA} with that of the phage-dependent promoters an up to 13 times higher production of recombinant protein was observed [27, 44]. Nevertheless the optimized xylose-inducible promoter still provides the strongest system for recombinant protein production in *B. megaterium* [44].

7.2.1.3 Additional Promoters for *Bacillus megaterium*

Besides the well characterized xylose-inducible promoter and the phage promoters other promoters were studied for recombinant protein production. During the analysis of the exoproteome of *B. megaterium* exposed to different growth conditions, the addition of sucrose to the growth medium resulted in a strong secretion of a protein which could be identified as a levansucrase [45]. The promoter region (P_{sacB}) of the corresponding gene was introduced in a *Bacillus*-*E. coli*-shuttle vector replacing the xylose-inducible promoter [46]. In contrast to P_{xylA}, P_{sacB} showed a basal expression already in the absence of its inducer sucrose. At the level of produced recombinant protein the native P_{xylA} was found to be twice as strong as P_{sacB} [46].

Additionally, a starch-inducible promoter P_{amyL} from *Bacillus amyloliquefaciens* was shown to function in *B. megaterium*. After the addition of starch to the growth medium, the recombinant production of an extracellular keratinase resulted in yields almost comparable with these achieved by using the native xylose-inducible P_{xylA} [23].

Finally, also the use of constitutive promoters for the recombinant plasmid-derived protein production was described. For the overproduction of two glucose dehydrogenases the native promoters of their corresponding genes were used. Both proteins were constitutively produced but in dramatically different amounts [22].

7.2.2 Production of Recombinant Proteins and Protein Complexes

7.2.2.1 Intracellular Protein Production

For *B. megaterium* the strength of most of the above mentioned promoters was quantified by using an engineered green fluorescent protein (GFP+, here referred to as GFP) as a model protein [48]. The amount of intracellularly produced GFP can spectroscopically be followed online during the whole recombinant production process. Maximal GFP titers of 1.25 g per liter during high cell density cultivations and yields of 82.5 mg per g of cell dry weight during shake flask cultivations were achieved using the optimized xylose-inducible promoter (Table 7.1, Fig. 7.2 I) [40]. Especially for the coproduction of different proteins belonging to larger protein complexes also weak production of the recombinant target protein resulting from less active promoters is of interest (Table 7.1). This results in a reduced metabolic burden to the cell [32]. The feasible production of recombinant GFP encoded on two different compatible vectors individually controlled by the native P_{xylA} is of relevance for the production of multiple proteins and protein complexes [26, 27].

With a size of 26.9 kDa, the model protein GFP is rather small [48]. However, *B. megaterium* showed the potential to recombinantly produce proteins with molecular weights of more than 300,000 Da. Two *Clostridium difficile* toxins, toxin A and toxin B (rTcdA and rTcdB), with molecular weights of 308,000 Da and 280,000 Da, respectively, were recombinantly overproduced in *B. megaterium* (Table 7.2). These recombinant proteins revealed identical biological activities as their native counterparts isolated from *C. difficile* which was not observed when recombinantly producing these proteins in *E. coli* [49].

Further on, examples exist showing that *B. megaterium* can be used for whole-cell bioconversion while recombinantly producing at least 2 or even up to 14 proteins in parallel. For the biotransformation of D-fructose into D-mannitol the mannitol dehydrogenase from *Leuconostoc pseudomesenteroides* and the formate dehydrogenase from *Mycobacterium vaccae*, latter used for generating the NADH reduction equivalents, were recombinantly coproduced in *B. megaterium*. The coexpression of the corresponding recombinant genes was plasmid-driven under the control of the xylose-inducible promoter. Significant D-mannitol production was observed [50]. Another example for a whole-cell bioconversion using the *B. megaterium* system was the hydroxylation of the pentacyclic triterpene 11-keto-β-boswellic acid (KBD) in the presence of a recombinant cytochrome P450 system [51, 52]. A recombinant P450 was coproduced with up to two redox partners plasmid-encoded in one operon controlled by the xylose-inducible promoter. The different combinations of redox partners responsible for electron transfer to the cytochrome P450 resulted in differential product formation indicating the importance of suitable protein interaction and redox partners [52].

For *in vivo* analysis of vitamin B_{12} (cobalamin) production in its natural producer *B. megaterium* the overexpression of 14 genes representing the whole *cobI*-operon of *B. megaterium* responsible for vitamin B_{12} production was performed vector-derived under the control of the constitutive promoter P_{cbi} and the xylose-inducible promoter P_{xylA}, respectively [53, 54]. All proteins were found being overproduced after the addition of xylose. The xylose-induced overproduction resulted in a 5.5 higher amount of

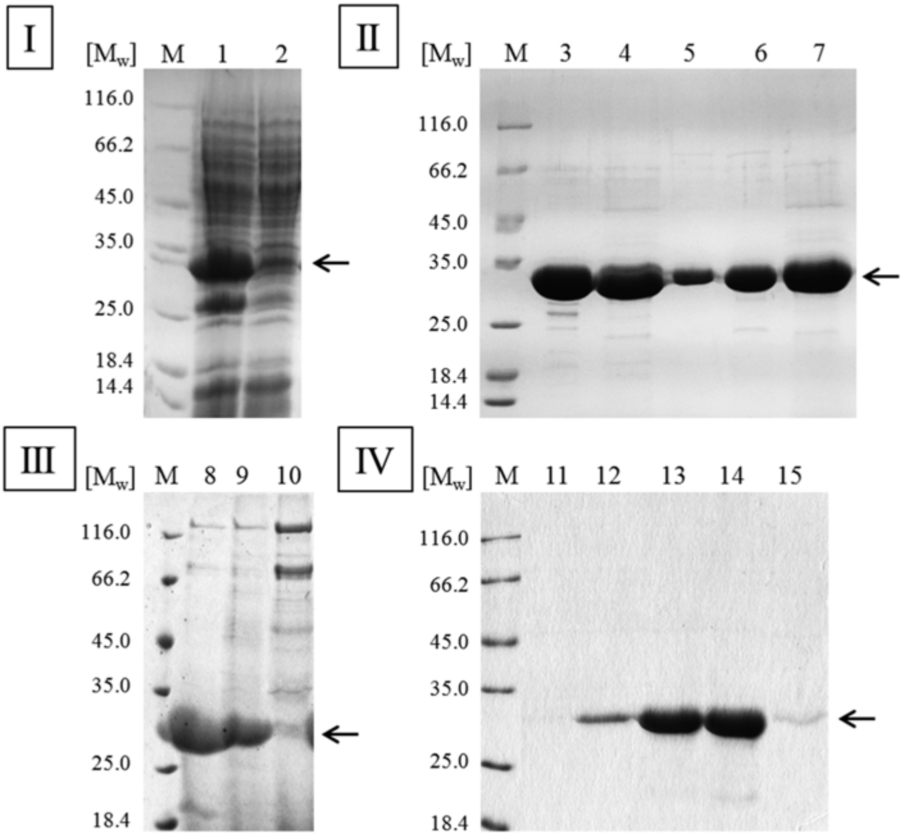


Fig. 7.2 SDS-PAGE analyses of proteins produced and secreted by recombinant *Bacillus megaterium*. (*I*) Intracellular proteins of recombinant *B. megaterium* plasmid-carrying strains. The green fluorescent protein (GFP) was overproduced under the control of the optimized xylose-inducible promoter. Lane 1: GFP-overproduction after the addition of xylose; lane 2: control without addition of xylose; *arrow* indicates molecular weight of GFP. (*II*) Elution fractions of five affinity purifications of recombinant GFP produced by different *B. megaterium* plasmid-carrying strains. Lane 3: GFP-StrepII; lane 4: GFP-His$_6$; lane 5: StrepII-GFP; lane 6: StrepII-GFP; lane 7: His-GFP; *arrow* indicates molecular weight of GFP. (*III*) Extracellular proteins of recombinant *B. megaterium* plasmid-carrying strains (corresponding to 1.5 ml of cell-free growth medium). Protein A from *Staphylococcus aureus* fused to the signal peptide of the lipase A (lane 8) or to its native signal peptide (lane 9) was overproduced and secreted controlled by the optimized xylose-inducible promoter. Lane 10: control without addition of xylose; *arrow* indicates molecular weight of Protein A. (*IV*) Different elution fractions (lanes 11–15) of an affinity purification of recombinant His$_6$-tagged protein A from cell-free growth medium of recombinant *B. megaterium* plasmid-carrying strains; *arrow* indicates molecular weight of Protein A. Lane M: Unstained Protein Molecular Weight Markers (Thermo Fisher)

cobyric acid, a precursor of the vitamin B$_{12}$, compared to the overproduction using the constitutive system [53]. This clearly indicates the capacity of this organism to coproduce 14 functional enzymes in adequate amounts. Similar observation was made for the overproduction of the enzymes encoded by the *hemAXCDBL* operon responsible for the production of the tetrapyrrole precursor molecule uroporphyrinogen III. The genomic integration of the xylose-inducible promoter P$_{xylA}$ upstream of this operon also enhanced production of the tetrapyrrole vitamin B$_{12}$ in *B. megaterium* [55].

7.2.2.2 Extracellular Protein Production

The secretion of recombinant proteins can drastically reduce the time, effort and cost for their

Table 7.2 Recombinant proteins produced and secreted with *Bacillus megaterium*

Protein	Source	Promoter	SP	Maximal product titer	References
Intracellular					
GFP	*Aequorea victoria*	P_{xylA}, P_{T7}, P_{sacB}, P_{K1E}, P_{SP6}	n/a	1.25 g/L	(Table 7.1)
Toxin A and B	*Clostridium difficile*	P_{xylA}	n/a	10 mg/L (purified)	[49]
Individual vitamin B_{12} biosynthesis enzymes	*Bacillus megaterium*	P_{xylA}	n/a	100 mg/L (purified)	[74, 75]
14 vitamin B_{12} biosynthesis enzymes (multicistronic)	*Bacillus megaterium*	P_{xylA}, P_{cbi}	n/a	n. d.	[53]
Formate dehydrogenase and mannitol dehydrogenase (bicistronic)	*Mycobacterium vaccae* and *Leuconostoc pseudomesenteroides*	P_{xylA}	n/a	n. d.	[50]
Cytochrome P450, AdR and Adx (tricistronic)	*Bacillus megaterium*; rind	P_{xylA}	n/a	n. d.	[51]
Cytochrome P450 and redox partners (tricistronic)	*Bacillus megaterium*	P_{xylA}	n/a	n. d.	[52]
Extracellular					
Levansucrase	*Lactobacillus reuteri*	P_{xylA}, P_{T7}	SP_{lipA}	4 mg/L	[27, 62, 63]
Levensucrase	*Bacillus megaterium*	P_{xylA}, P_{sacB}	SP_{sacB}	0.5 g/L	[6, 46]
Antibody fragment D1.3scFv	mouse	P_{xylA}	SP_{lipA}	12 mg/L	[70, 72]
Dextransucrase	*Leuconostoc mesenteroides*	P_{xylA}	SP_{native}	28,600 U/L	[38, 61]
Hydrolase	*Thermobifida fusca*	P_{xylA}, P_{T7}	SP_{vpr}, SP_{nprM}, SP_{lipA}, SP_{pga}, SP_{sacB}, SP_{asp}	7,200 U/L	[27, 40, 76]
Penicillin G Amidase	*Bacillus megaterium*	P_{xylA}	SP_{pga}, SP_{lipA}	40 mg/L	[67]
Toxin B	*Clostridium difficile*	P_{xylA}	SP_{lipA}	n. d.	[49]
Endoglucanase and -glucosidase (coproduction of individual / fused proteins)	*Bacillus amyloliquefaciens*	P_{xylA}, P_{amyL}	SP_{lipA}, SP_{native}	n. d.	[8]
Keratinase	*Bacillus licheniformis*	P_{xylA}, P_{amyL}	SP_{native}	186 U/mL	[23, 65]

Promoters, signal peptides (SP) and target genes are located on *Bacillus-E.coli*-shuttle vectors

AdR bovine adrenodoxin reductase, *Adx* bovine adrenodoxin, P_{xylA} xylose-inducible promoter from *B. megaterium*, P_{T7} T7-RNA-polymerase dependent promoter from the phage T7, P_{sacB} sucrose-inducible promoter from *B. megaterium*, P_{K1E} K1E-RNA-polymerase dependent promoter from the *E. coli* phage K1E, P_{SP6} SP6-RNA-polymerase dependent promoter from the *Salmonella typhimurium* phage SP6, P_{amyL} starch-inducible promoter from *B. amyloliquefaciens*, P_{cbi} constitutive promoter of *cbi* operon of *B. megaterium*, *SP* signal peptide, SP_{native} native signal peptide of corresponding protein, SP_{lipA} signal peptide of lipase A, SP_{pga} signal peptide of penicillin G amidase, SP_{sacB} signal peptide of levansucrase, SP_{nprM} signal peptide of the neutral protease, SP_{vpr} signal peptide of minor extracellular protease, SP_{asp} signal peptide of computationally designed artificial signal peptide, *n.d.* not determined, *n/a* not analyzed purification as the protein of interest can directly be isolated from the cell-free broth [2]. Moreover, continuous production processes can be realized in this way. As a Gram-positive organism, *B. megaterium* is able to secrete proteins directly to the growth medium using amongst others the so called SEC (secretion)-dependent pathway which was intensively analyzed for Bacilli [56]. About 90 % of the secreted proteins in Bacilli are transported by this route in an unfolded state directed by an N-terminally localized signal peptide. Such SEC-dependent signal peptides consist of an N-terminal positively charged region (N-region) with up to 11 basic amino acids like arginine or lysine, followed by a hydrophobic core region (H-region) containing a helix-breaking residue which is mostly a glycine or a proline residue and a more hydrophilic C-terminal region (C-region). The C-region ends with the signal peptidase cleavage site [57].

An alternative secretion pathway is the so called TAT (Twin Arginine Transport) system described for the Gram-positive *B. subtilis* and some others [58]. Via this pathway proteins are transported in a completely folded, active conformation. The proteins responsible for the pore formation as well as a homologue to a TAT-dependent protein described for *B. subtilis*, the alkaline phosphatase PhoD, were also found in *B. megaterium* [9, 58]. This pathway can be used for the transport of proteins already binding their corresponding cofactor and, as already shown for *E. coli*, it is also suitable for the transport of protein complexes already assembled in the cytoplasm [59].

As the SEC-pathway presents the dominant route for protein secretion in *B. megaterium* [9], major focus was laid on this route for its biotechnological utilization. Different signal peptides have been deduced from the genome sequence involved in the SEC-dependent protein secretion [9, 40, 60]. The first protein recombinantly produced and secreted in an active form was the 170 kDa dextransucrase DsrS from *Leuconostoc mesenteroides*. In a high cell density cultivation reaching 80 g of cell dry weight per liter a volumetric activity of 28,600 Units per liter of this large recombinant enzyme was measured [38]. The signal peptidase SipM, required for the cleavage of the signal peptide after protein secretion, was coproduced with the DsrS to enhance protein secretion in *B. megaterium*, controlled by its own native, constitutive promoter. The secretion of DsrS was enhanced almost 4 times via the gentle coproduction of SipM [61]. A similar effect of *sipM* coexpression could be shown for the secretion of a recombinant levansucrase using *B. megaterium* [62]. The secretion of recombinant *C. difficile* toxin B (rTcdB) with a molecular weight of 280,000 Da demonstrates that even larger proteins can be recombinantly produced and secreted by *B. megaterium* [49].

Once proteins targeted to be secreted are getting smaller their yield can be increased. Furthermore, the promoter strength, the origin of the target protein, the codon usage of the target gene, the employed secretion signal and the culture conditions turned out to be of great importance [29–31]. While the secretion of a *Lactobacillus reuteri* levansucrase with a molecular weight of 110,000 Da was in the low mg per liter range [62, 63], the recombinant secretion of the 51 kDa *B. megaterium* levansucrase reached more than 0.5 g per liter cell culture being the dominant protein in the culture supernatant corresponding to around 70 % of all secreted proteins [6, 45]. Similar observations were made for Protein A from *Staphylococcus aureus* recombinantly produced and secreted by *B. megaterium* (Fig. 7.2 III).

For the quantitative comparison of intracellular production of recombinant proteins in *E. coli* with recombinant production and secretion by *B. megaterium* a recombinant β-mannanase was used [64]. The protein was produced by both organisms in about the same amount. Anyway, while the protein remained intracellularly in *E. coli*, *B. megaterium* secreted the protein into the growth medium yielding already very pure protein [64]. For a keratinase from *B. licheniformis* the production and subsequent secretion with *B. megaterium* was found to yield more recombinant protein in the growth medium compared to the intracellular production in *E. coli* [65].

A really interesting approach to secrete two or more interacting proteins using the SEC-dependent pathway is the fusion of their coding sequences already at the DNA level. This was done for the secretion of an endoglucanase and an endoglucosidase, both used for the hydrolyzation of sugarcane bagasse. Here, the authors fused the proteins at the level of DNA via a sequence for a flexible structured and water soluble five amino acid linker region and further on with the signal peptide of the *B. megaterium* lipase A at the N-terminus of the fusion construct [8, 63]. For comparison, the proteins were also coproduced individually equipped with the N-terminal secretion signal of the lipase A for the endoglucosidase and the native signal peptide for the endoglucanase. An identical activity for the fusion enzyme was detected when compared to the activity of the coproduced individual proteins mixed in equal amounts [8].

The penicillin G amidase (PGA) provides an example of a protein which is processed into its two subunits after export [66]. *B. megaterium* PGA could be recombinantly produced and secreted by *B. megaterium* reaching 40 mg/L in the growth medium [67]. Interestingly the secretion was enhanced when replacing the natural secretion signal of the *B. megaterium* PGA with that of the *B. megaterium* lipase A demonstrating that efficiency of protein secretion depends on the nature of the signal peptide in combination with the physical properties of the target protein. Hence, the choice of the optimal suited signal peptide for secreting a given recombinant protein is of high importance [66–69]. Consequently, a set of *B. megaterium* signal peptides, predicted to be very effective, were successfully tested and made available for protein export approaches. These include signal peptides from the lipase A (*lipA*), the penicillin G amidase (*pga*), the levansucrase (*sacB*), the neutral protease (*nprM*), the minor extracellular protease (*vpr*) as well as a computationally designed artificial signal peptide (*asp*) [40, 46, 62, 63, 67]. After introducing their coding sequences individually upstream of the MCS in the *B. megaterium* expression system, cloning of the gene(s) of interest is possible [40].

Further, the production and secretion of proteins naturally not occurring in prokaryotes is another big challenge for bacterial systems. For *B. megaterium* it was shown that antibody fragments of different sizes and structures can be secreted via the SEC-pathway even at the 100 l scale [70–72]. These fragments namely single chain fragment variable (scFv) and single chain fragment antigen binding (scFab) have molecular weights of around 27,000 Da and 51,000 Da, respectively [73]. They consist of two fused domains representing parts of the light and the heavy chain of a natural antibody. As these domains naturally are not connected, they are linked via a chain of 15–34 amino acids. These fusion proteins were produced and finally secreted using the SEC-pathway directed via the N-terminal secretion signal of the lipase A by *B. megaterium* in amounts of up to 12 mg per liter (Table 7.2) [31].

7.2.3 Purification of Recombinant Proteins Produced with *Bacillus megaterium*

The fast, simple and cheap production and purification of recombinant proteins in an apparently pure form has become a major task of biotechnological research and industry. With *E. coli* as expression host, fast and simple purification of recombinant proteins can be achieved using so called affinity tags. Numerous established vector systems are commercially available for this production host [29, 77]. Similarly, DNA sequences encoding small affinity tags (His$_6$-tag, StrepII-tag) have also been introduced in the *B. megaterium* plasmid system upstream or downstream of the MCS generating recombinant proteins carrying an N- or an C-terminal His$_6$- or StrepII-tag [40, 47, 78]. To remove these tags specific cleavage sites for tag removal by the protease factor X$_a$ or the tobacco etch virus (TEV) protease were additionally introduced between MCS and N-terminal tag [79, 80].

These systems were successfully used for the production and purification of ten different highly oxygen-sensitive enzymes involved in vitamin B_{12}-biosynthesis in *B. megaterium* via N- or C-terminally fused His_6-tags. The whole purification process occurred under oxygen-free conditions after which all of these enzymes showed their corresponding activity. Some of them were also found to occur in dimeric or even tetrameric state after purification with yields of up to 100 mg/L [74, 75]. The same was shown for the production and purification of the membrane associated heme biosynthesis enzyme HemG fused to an His_6-tag using *B. megaterium* [81]. The StrepII-tag was successfully employed for the purification of recombinant GFP resulting in apparently pure target protein (Fig. 7.2 II). In summary using these systems up to 70 % of the recombinant protein can be purified after cell disruption (Fig. 7.2 I+II) [47, 75].

Further on the small purification tags (His_6-tag, StrepII-tag) were combined with the signal peptide of the lipase A responsible for the secretion of a recombinant target protein using the *B. megaterium* vector system [62, 63]. When these systems were employed for the recombinant overproduction and secretion of target proteins like the levansucrases of *L. reuteri*, the hydrolase of *Thermobifida fusca* or Protein A from *S. aureus* they became the dominant protein in the cell free growth medium (Fig. 7.2 III) [40, 62, 76]. Apparently pure and active protein directly purified from the cell free medium was achieved after one single purification step without any changes of pH or salt concentrations (Fig. 7.2 IV) [62, 63]. The purification process can also be integrated into the fermentation process using purification material like functionalized magnetic beads which are easy to separate from the culture broth [62, 82, 83].

These different purification systems are also suitable for the production and purification of proteins interacting or forming complexes. For this, vectors equipped with DNA sequences of purification tags fused to one protein of the complex can be combined with vectors lacking such sequences encoding the complex partners. This will result in the copurification of target proteins belonging to one multi-subunit protein complex (Fig. 7.3).

7.3 Omics-Driven Analyses of the Recombinant Protein Production Process of *Bacillus megaterium*

7.3.1 Systems Biology of *Bacillus megaterium* for the optimization of Recombinant Protein Production

Already in 2005/2006 first approaches using 2D-gel based proteomics and initial fluxomic analyses were undertaken to compare *B. megaterium* strains recombinantly producing and secreting the dextransucrase DsrS or the hydrolase of *T. fusca* TFH with non-producing *B. megaterium* strains [84–88]. However, due to technical limitations and a missing genome sequence of *B. megaterium* the success was limited. Nevertheless, it was the first time that significant differences in the metabolic flux of *B. megaterium* recombinantly producing and secreting the TFH using different carbon sources were observed and quantified [88].

The genome sequence of *B. megaterium* published in 2011 opened the door for a systems biology access to the protein production and secretion by this biotechnologically relevant organism [9, 24, 26]. Numerous protocols for transcriptome, proteome, metabolome and fluxome analyses were developed, published and are currently utilized [24, 26, 31]. First results indicated significant differences of the gene regulating networks of *B. megaterium* compared to that of *B. subtilis* and *B. licheniformis* [89, 90]. A map of the metabolic network of *B. megaterium* was deduced from its genome sequence and provided the base for flux balance analyses [87, 88, 91].

7.3.2 Genetic Engineering of *Bacillus megaterium*

The so called omics technologies such as transcriptomics, proteomics, metabolomics and fluxomics provide molecular insights into processes of a living organism. They can indicate

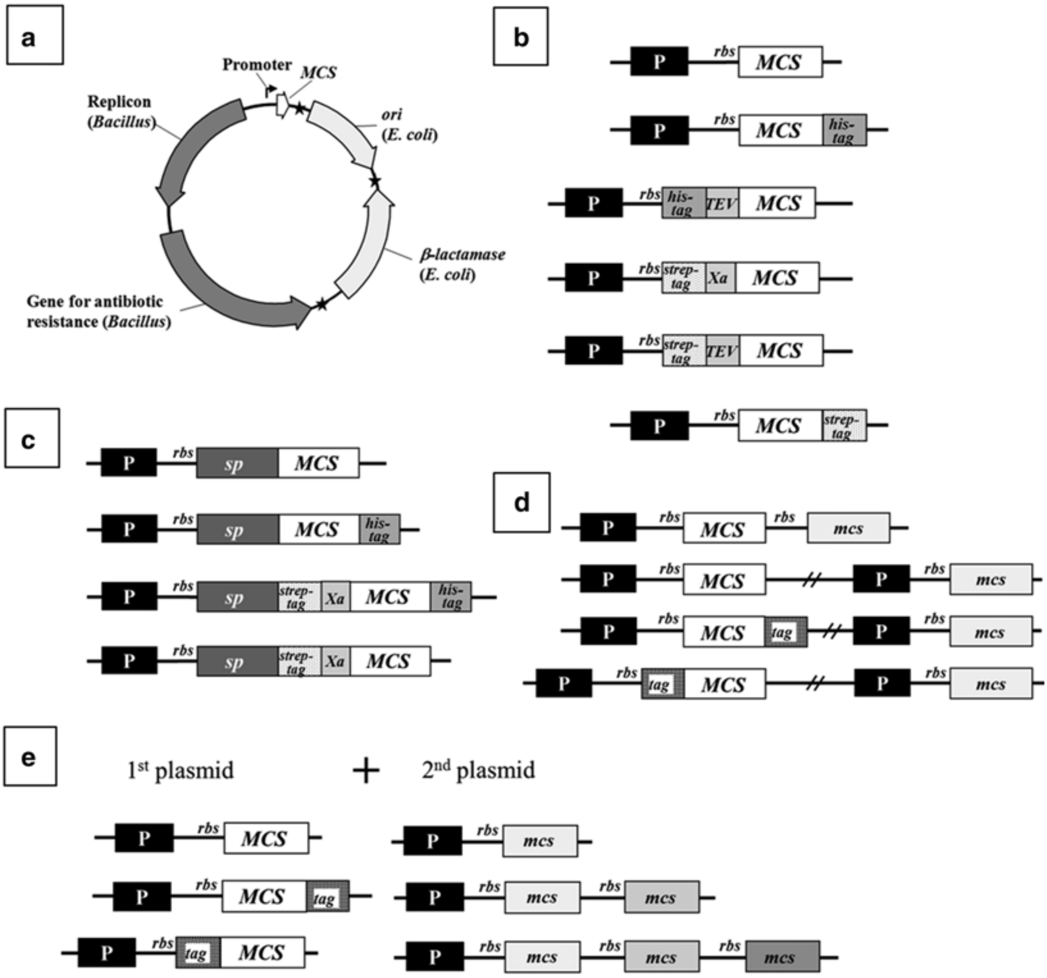


Fig. 7.3 Schemes of *Bacillus megaterium* vectors for the production, secretion and purification of recombinant proteins. (**a**) *Bacillus*-*E. coli*-shuttle vector for cloning in *E. coli* (*light gray* – functional elements for *E. coli*: origin of replication (*ori*), gene encoding β-Lactamase) and recombinant protein production in *B. megaterium* (*dark gray* – functional elements for *B. megaterium*: replicon (belonging to different compatibility classes: *repU*/*oriU*, *repM100*, *repM700*), gene for antibiotic resistance (tetracycline, erythromycin, chloramphenicol)). Further variable elements are the promoter (Table 7.2) and the multiple cloning site (MCS). The stars indicate possible positions for introduction of genes for coexpression. (**b–e**) Detailed description of the variable elements indicated "promoter" and "MCS" in (**a**). (**b**) Vector constructs for the purification of recombinant intracellular protein with integrated His$_6$- or StrepII-tag. The N-terminal located tags can be cleaved off using the protease factor X$_a$ (Xa) or the tobacco etch virus (TEV) protease [2, 26, 40, 46, 47]. (**c**) Vector constructs for the secretion and purification of recombinant proteins. The protein is guided via a specific N-terminally localized signal peptide to the growth medium (Table 7.2). Integrated His$_6$- or StrepII-tags allow the purification of the secreted protein. The N-terminally localized StrepII-tag can be cleaved off by the protease factor X$_a$ (Xa) [40, 62, 63]. (**d**) Vector constructs for the coexpression of two or more different genes. One gene can be fused to a purification tag (His$_6$- or StrepII-tag). Further genes can be placed in an operon-like structure downstream of the first gene or combined with a promoter at any position indicated with stars in (**a**). (**e**) Vector constructs for the coexpression of two or more different genes. One gene can be fused to a purification tag (His$_6$- or StrepII-tag). Further genes can be placed on a second plasmid under control of a second promoter as single gene or in multicistronic manner. *rbs* ribosome binding site, *MCS*/*mcs* multiple cloning site, *P* promoter

limitations during the production process of a recombinant protein which could be overcome by the coproduction of required additional proteins or by the knock-out of inhibiting genes. Many genetic tools and different methods are available for the genetic engineering of *B. megaterium* [24]. Using these methods, multiple mutant strains were constructed within the last years. Strains showing a higher stability of secreted proteins by the deletion of the main extracellular protease were achieved [92]. Also new strains which do not metabolize the inducer xylose for the inducible promoter P_{xylA} anymore were constructed and successfully tested indicating their use in the recombinant protein production process [36, 67]. For industrial applications the biosafety of the used organism is of high impact. Here, *B. megaterium* strains with a high sensitivity to UV light, deficient in sporulation and genetic recombination were constructed [11, 12, 93]. Now, with the combination of the data from different omics technologies based on the genome sequence and genetic modification the construction of further new *B. megaterium* strains is in progress [24, 26, 31].

7.4 Future of *Bacillus megaterium* for the Production of Multi-subunit Protein Complexes

We only started the recombinant production of multi-subunit protein complexes using *B. megaterium*. Nevertheless, useful multiple vector systems are already available harboring inducible and constitutive promoters of different and adjustable strength with enhanced multiple cloning sites allowing parallel cloning and also encoding different affinity tags for purification of recombinant proteins or protein complexes (Fig. 7.3). Most importantly, compatible origins of replication for using two or even more individual recombinant plasmids provide the basis for the production for multi-subunit protein complexes. A variety of different signal peptides for the secretion of recombinant proteins complete the picture (Fig. 7.3 C and Table 7.2). Further, high molecular weight protein production does not constitute problems for *B. megaterium* [38, 49, 61]. First examples have already shown that this organism is able to recombinantly overproduce up to 14 proteins in parallel in an active form [53]. This clearly indicates that *B. megaterium* represents an ideal host for the coproduction of proteins as well as the production of multi-subunit protein complexes.

Acknowledgments This work was financially supported within the framework of the Collaborative Research Center SFB 578 “From Gene to Product” and the priority project SPP1617 “Phenotypic Heterogeneity and Sociobiology of Bacterial Populations” by the German Research Foundation (DFG). The author would like to kindly thank Dieter Jahn (Technische Universität Braunschweig, Germany) and many active and former members of his team for support and fruitful discussions, Manfred Rohde from the Helmholtz Centre for Infection Research (Braunschweig, Germany) for electron microscopic pictures of *B. megaterium*, the MoBiTec GmbH (Göttingen, Germany) for support and collaboration and Friedhelm Meinhardt and his team (Münster University, Germany) for support, fruitful discussions and collaboration.

References

1. De Bary A (1884) Vergleichende Morphologie und Biologie der Pilze, Mycetozoen und Bacterien. Wilhelm Engelmann, Leipzig
2. Vary PS, Biedendieck R, Fuerch T, Meinhardt F, Rohde M, Deckwer WD, Jahn D (2007) *Bacillus megaterium* – from simple soil bacterium to industrial protein production host. Appl Microbiol Biotechnol 76(5):957–967
3. Vary P (1992) Development of genetic engineering in *Bacillus megaterium*. Biotechnology 22:251–310
4. Vary PS (1994) Prime time for *Bacillus megaterium*. Microbiology 140(Pt 5):1001–1013
5. Xu H, Qin Y, Huang Z, Liu Z (2014) Characterization and site-directed mutagenesis of an alpha-galactosidase from the deep-sea bacterium *Bacillus megaterium*. Enzym Microb Technol 56:46–52
6. Korneli C, Biedendieck R, David F, Jahn D, Wittmann C (2013) High yield production of extracellular recombinant levansucrase by *Bacillus megaterium*. Appl Microbiol Biotechnol 97(8):3343–3353
7. Zheng H, Liu Y, Liu X, Han Y, Wang J, Lu F (2012) Overexpression of a *Paenibacillus campinasensis* xylanase in *Bacillus megaterium* and its applications to biobleaching of cotton stalk pulp and saccharification of recycled paper sludge. Bioresour Technol 125:182–187

8. Kurniasih SD, Alfi A, Natalia D, Radjasa OK, Nurachman Z (2014) Construction of individual, fused, and co-expressed proteins of endoglucanase and beta-glucosidase for hydrolyzing sugarcane bagasse. Microbiol Res 169(9–10):725–732
9. Eppinger M, Bunk B, Johns MA, Edirisinghe JN, Kutumbaka KK, Koenig SS, Creasy HH, Rosovitz MJ, Riley DR, Daugherty S, Martin M, Elbourne LD, Paulsen I, Biedendieck R, Braun C, Grayburn S, Dhingra S, Lukyanchuk V, Ball B, Ul-Qamar R, Seibel J, Bremer E, Jahn D, Ravel J, Vary PS (2011) Genome sequences of the biotechnologically important *Bacillus megaterium* strains QM B1551 and DSM319. J Bacteriol 193(16):4199–4213
10. Liu L, Li Y, Zhang J, Zou W, Zhou Z, Liu J, Li X, Wang L, Chen J (2011) Complete genome sequence of the industrial strain *Bacillus megaterium* WSH-002. J Bacteriol 193(22):6389–6390
11. Nahrstedt H, Meinhardt F (2004) Structural and functional characterization of the *Bacillus megaterium uvrBA* locus and generation of UV-sensitive mutants. Appl Microbiol Biotechnol 65(2):193–199
12. Nahrstedt H, Schroder C, Meinhardt F (2005) Evidence for two *recA* genes mediating DNA repair in *Bacillus megaterium*. Microbiology 151(Pt 3):775–787
13. Meinhardt F, Busskamp M, Wittchen KD (1994) Cloning and sequencing of the *leuC* and *nprM* genes and a putative spo IV gene from *Bacillus megaterium* DSM319. Appl Microbiol Biotechnol 41(3):344–351
14. Couturier M, Bex F, Bergquist PL, Maas WK (1988) Identification and classification of bacterial plasmids. Microbiol Rev 52(3):375–395
15. Janniere L, Gruss A, Ehrlich S (1993) Plasmids. In: Sonenshein A, Hoch J, Losick R (eds) *Bacillus subtilis* and other gram-positive bacteria. American Society of Microbiology, Washington, DC, pp 625–645
16. Kieselburg MK, Weickert M, Vary PS (1984) Analysis of resident and transformant plasmids in *Bacillus megaterium*. Biotechnology 2:254–259
17. Stevenson DM, Kunnimalaiyaan M, Müller K, Vary PS (1998) Characterization of a theta plasmid replicon with homology to all four large plasmids of *Bacillus megaterium* QM B1551. Plasmid 40(3):175–189
18. Scholle MD, White CA, Kunnimalaiyaan M, Vary PS (2003) Sequencing and characterization of pBM400 from *Bacillus megaterium* QM B1551. Appl Environ Microbiol 69(11):6888–6898
19. Kunnimalaiyaan M, Vary PS (2005) Molecular characterization of plasmid pBM300 from *Bacillus megaterium* QM B1551. Appl Environ Microbiol 71(6):3068–3076
20. Sussman MD, Vary PS, Hartman C, Setlow P (1988) Integration and mapping of *Bacillus megaterium* genes which code for small, acid-soluble spore proteins and their protease. J Bacteriol 170(10):4942–4945
21. Rygus T, Hillen W (1991) Inducible high-level expression of heterologous genes in *Bacillus megaterium* using the regulatory elements of the xylose-utilization operon. Appl Microbiol Biotechnol 35(5):594–599
22. Meinhardt F, Stahl U, Ebeling W (1989) Highly efficient expression of homologous and heterologous genes in *Bacillus megaterium*. Appl Microbiol Biotechnol 30:343–350
23. Radha S, Gunasekaran P (2008) Sustained expression of keratinase gene under P_{xylA} and P_{amyL} promoters in the recombinant *Bacillus megaterium* MS941. Bioresour Technol 99(13):5528–5537
24. Biedendieck R, Borgmeier C, Bunk B, Stammen S, Scherling C, Meinhardt F, Wittmann C, Jahn D (2011) Systems biology of recombinant protein production using *Bacillus megaterium*. Methods Enzymol 500:165–195
25. Bernhard K, Schrempf H, Goebel W (1978) Bacteriocin and antibiotic resistance plasmids in *Bacillus cereus* and *Bacillus subtilis*. J Bacteriol 133(2):897–903
26. Biedendieck R, Bunk B, Furch T, Franco-Lara E, Jahn M, Jahn D (2010) Systems biology of recombinant protein production in *Bacillus megaterium*. Adv Biochem Eng Biotechnol 120:133–161
27. Gamer M, Fröde D, Biedendieck R, Stammen S, Jahn D (2009) A T7 RNA polymerase-dependent gene expression system for *Bacillus megaterium*. Appl Microbiol Biotechnol 82(6):1195–1203
28. Kunnimalaiyaan M, Stevenson DM, Zhou Y, Vary PS (2001) Analysis of the replicon region and identification of an rRNA operon on pBM400 of *Bacillus megaterium* QM B1551. Mol Microbiol 39(4):1010–1021
29. Terpe K (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol 72(2):211–222
30. Fernandez FJ, Vega MC (2013) Technologies to keep an eye on: alternative hosts for protein production in structural biology. Curr Opin Struct Biol 23(3):365–373
31. Korneli C, David F, Biedendieck R, Jahn D, Wittmann C (2013) Getting the big beast to work – systems biotechnology of Bacillus megaterium for novel high-value proteins. J Biotechnol 163(2):87–96
32. Glick BR (1995) Metabolic load and heterologous gene expression. Biotechnol Adv 13(2):247–261
33. Rygus T, Scheler A, Allmansberger R, Hillen W (1991) Molecular cloning, structure, promoters and regulatory elements for transcription of the *Bacillus megaterium* encoded regulon for xylose utilization. Arch Microbiol 155(6):535–542
34. Dahl MK, Degenkolb J, Hillen W (1994) Transcription of the xyl operon is controlled in *Bacillus subtilis* by tandem overlapping operators spaced by four base-pairs. J Mol Biol 243(3):413–424
35. Gärtner D, Degenkolb J, Ripperger JAE, Allmansberger R, Hillen W (1992) Regulation of the *B. subtilis* W23 xylose utilization operon: interaction

of Xyl repressor with xyl operator and the inducer xylose. Mol Gen Genet 232(3):415–422

36. Rygus T, Hillen W (1992) Catabolite repression of the xyl operon in *Bacillus megaterium*. J Bacteriol 174(9):3049–3055
37. Bolivar F, Rodriguez RL, Greene PJ, Betlach MC, Heyneker HL, Boyer HW, Crosa JH, Falkow S (1977) Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene 2(2):95–113
38. Malten M, Hollmann R, Deckwer WD, Jahn D (2005) Production and secretion of recombinant *Leuconostoc mesenteroides* dextransucrase DsrS in *Bacillus megaterium*. Biotechnol Bioeng 89(2):206–218
39. Atanassov I, Stefanova K, Tomova I, Kamburova M (2013) Seamless GFP and GFP-amylase cloning in Gateway shuttle vector, expression of the recombinant proteins in *E. coli* and *Bacillus megaterium* and assessment the GFP-amylase thermostability. Biotechnol Biotechnol Equip 27:4172–4180
40. Stammen S, Müller BK, Korneli C, Biedendieck R, Gamer M, Franco-Lara E, Jahn D (2010) High-yield intra- and extracellular protein production using *Bacillus megaterium*. Appl Environ Microbiol 76(12):4037–4046
41. Tabor S, Richardson CC (1985) A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes. Proc Natl Acad Sci U S A 82(4):1074–1078
42. Brunschwig E, Darzins A (1992) A two-component T7 system for the overexpression of genes in *Pseudomonas aeruginosa*. Gene 111(1):35–41
43. Conrad B, Savchenko RS, Breves R, Hofemeister J (1996) A T7 promoter-specific, inducible protein expression system for *Bacillus subtilis*. Mol Gen Genet 250(2):230–236
44. Stammen S, Schuller F, Dietrich S, Gamer M, Biedendieck R, Jahn D (2010) Application of *Escherichia coli* phage K1E DNA-dependent RNA polymerase for i*n vitro* RNA synthesis and *in vivo* protein production in *Bacillus megaterium*. Appl Microbiol Biotechnol 88(2):529–539
45. Homann A, Biedendieck R, Gotze S, Jahn D, Seibel J (2007) Insights into polymer versus oligosaccharide synthesis: mutagenesis and mechanistic studies of a novel levansucrase from *Bacillus megaterium*. Biochem J 407(2):189–198
46. Biedendieck R, Gamer M, Jaensch L, Meyer S, Rohde M, Deckwer WD, Jahn D (2007) A sucrose-inducible promoter system for the intra- and extracellular protein production in *Bacillus megaterium*. J Biotechnol 132(4):426–430
47. Biedendieck R, Yang Y, Deckwer WD, Malten M, Jahn D (2007) Plasmid system for the intracellular production and purification of affinity-tagged proteins in *Bacillus megaterium*. Biotechnol Bioeng 96(3):525–537
48. Kaltwasser M, Wiegert T, Schumann W (2002) Construction and application of epitope- and green fluorescent protein-tagging integration vectors for Bacillus subtilis. Appl Environ Microbiol 68(5):2624–2628
49. Yang G, Zhou B, Wang J, He X, Sun X, Nie W, Tzipori S, Feng H (2008) Expression of recombinant *Clostridium difficile* toxin A and B in *Bacillus megaterium*. BMC Microbiol 8:192
50. Bäumchen C, Roth AH, Biedendieck R, Malten M, Follmann M, Sahm H, Bringer-Meyer S, Jahn D (2007) D-mannitol production by resting state whole cell biotrans-formation of D-fructose by heterologous mannitol and formate dehydrogenase gene expression in *Bacillus megaterium*. Biotechnol J 2(11):1408–1416
51. Bleif S, Hannemann F, Zapp J, Hartmann D, Jauch J, Bernhardt R (2012) A new *Bacillus megaterium* whole-cell catalyst for the hydroxylation of the pentacyclic triterpene 11-keto-beta-boswellic acid (KBA) based on a recombinant cytochrome P450 system. Appl Microbiol Biotechnol 93(3):1135–1146
52. Brill E, Hannemann F, Zapp J, Bruning G, Jauch J, Bernhardt R (2014) A new cytochrome P450 system from *Bacillus megaterium* DSM319 for the hydroxylation of 11-keto-beta-boswellic acid (KBA). Appl Microbiol Biotechnol 98(4):1701–1717
53. Moore SJ, Mayer MJ, Biedendieck R, Deery E, Warren MJ (2014) Towards a cell factory for vitamin B_{12} production in *Bacillus megaterium*: bypassing of the cobalamin riboswitch control elements. N Biotechnol 31(6):553–561
54. Raux E, Lanois A, Warren MJ, Rambach A, Thermes C (1998) Cobalamin (vitamin B_{12}) biosynthesis: identification and characterization of a *Bacillus megaterium cobI* operon. Biochem J 335(Pt 1):159–166
55. Biedendieck R, Malten M, Barg H, Bunk B, Martens JH, Deery E, Leech H, Warren MJ, Jahn D (2010) Metabolic engineering of cobalamin (vitamin B_{12}) production in *Bacillus megaterium*. Microb Biotechnol 3(1):24–37
56. Harwood CR, Cranenburgh R (2008) *Bacillus* protein secretion: an unfolding story. Trends Microbiol 16(2):73–79
57. Antelmann H, Tjalsma H, Voigt B, Ohlmeier S, Bron S, van Dijl JM, Hecker M (2001) A proteomic view on genome-based signal peptide predictions. Genome Res 11(9):1484–1502
58. Palmer T, Berks BC (2003) Moving folded proteins across the bacterial cell membrane. Microbiology 149(Pt 3):547–556
59. Sargent F (2007) The twin-arginine transport system: moving folded proteins across membranes. Biochem Soc Trans 35(Pt 5):835–847
60. Hiller K, Grote A, Scheer M, Munch R, Jahn D (2004) PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 32(Web Server issue):W375–W379
61. Malten M, Nahrstedt H, Meinhardt F, Jahn D (2005) Coexpression of the type I signal peptidase gene *sipM* increases recombinant protein production and export in *Bacillus megaterium* MS941. Biotechnol Bioeng 91(5):616–621

62. Biedendieck R, Beine R, Gamer M, Jordan E, Buchholz K, Seibel J, Dijkhuizen L, Malten M, Jahn D (2007) Export, purification, and activities of affinity tagged *Lactobacillus reuteri* levansucrase produced by *Bacillus megaterium*. Appl Microbiol Biotechnol 74(5):1062–1073
63. Malten M, Biedendieck R, Gamer M, Drews AC, Stammen S, Buchholz K, Dijkhuizen L, Jahn D (2006) A *Bacillus megaterium* plasmid system for the production, export, and one-step purification of affinity-tagged heterologous levansucrase from growth medium. Appl Environ Microbiol 72(2):1677–1679
64. Summpunn P, Chaijan S, Isarangkul D, Wiyakrutta S, Meevootisom V (2011) Characterization, gene cloning, and heterologous expression of beta-mannanase from a thermophilic *Bacillus subtilis*. J Microbiol 49(1):86–93
65. Radha S, Gunasekaran P (2007) Cloning and expression of keratinase gene in *Bacillus megaterium* and optimization of fermentation conditions for the production of keratinase by recombinant strain. J Appl Microbiol 103(4):1301–1310
66. Panbangred W, Weeradechapon K, Udomvaraphant S, Fujiyama K, Meevootisom V (2000) High expression of the penicillin G acylase gene (*pac*) from *Bacillus megaterium* UN1 in its own *pac* minus mutant. J Appl Microbiol 89(1):152–157
67. Yang Y, Biedendieck R, Wang W, Gamer M, Malten M, Jahn D, Deckwer WD (2006) High yield recombinant penicillin G amidase production and export into the growth medium using *Bacillus megaterium*. Microb Cell Factories 5:36
68. Ruiz C, Blanco A, Pastor FI, Diaz P (2002) Analysis of *Bacillus megaterium* lipolytic system and cloning of LipA, a novel subfamily I.4 bacterial lipase. FEMS Microbiol Lett 217(2):263–267
69. Brockmeier U, Caspers M, Freudl R, Jockwer A, Noll T, Eggert T (2006) Systematic screening of all signal peptides from *Bacillus subtilis*: a powerful strategy in optimizing heterologous protein secretion in Gram-positive bacteria. J Mol Biol 362(3):393–402
70. Jordan E, Hust M, Roth A, Biedendieck R, Schirrmann T, Jahn D, Dubel S (2007) Production of recombinant antibody fragments in *Bacillus megaterium*. Microb Cell Factories 6:2
71. Jordan E, Al-Halabi L, Schirrmann T, Hust M, Dubel S (2007) Production of single chain Fab (scFab) fragments in *Bacillus megaterium*. Microb Cell Factories 6:38
72. David F, Steinwand M, Hust M, Bohle K, Ross A, Dubel S, Franco-Lara E (2011) Antibody production in *Bacillus megaterium*: strategies and physiological implications of scaling from microtiter plates to industrial bioreactors. Biotechnol J 6(12):1516–1531
73. Hust M, Jostock T, Menzel C, Voedisch B, Mohr A, Brenneis M, Kirsch MI, Meier D, Dubel S (2007) Single chain Fab (scFab) fragment. BMC Biotechnol 7:14
74. Moore SJ, Biedendieck R, Lawrence AD, Deery E, Howard MJ, Rigby SE, Warren MJ (2013) Characterization of the enzyme CbiH$_{60}$ involved in anaerobic ring contraction of the cobalamin (vitamin B$_{12}$) biosynthetic pathway. J Biol Chem 288(1):297–305
75. Moore SJ, Lawrence AD, Biedendieck R, Deery E, Frank S, Howard MJ, Rigby SE, Warren MJ (2013) Elucidation of the anaerobic pathway for the corrin component of cobalamin (vitamin B$_{12}$). Proc Natl Acad Sci U S A 110(37):14906–14911
76. Yang Y, Malten M, Grote A, Jahn D, Deckwer WD (2007) Codon optimized *Thermobifida fusca* hydrolase secreted by *Bacillus megaterium*. Biotechnol Bioeng 96(4):780–794
77. Terpe K (2003) Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol 60(5):523–533
78. www.mobitec.de (2014) hp-vector expression systems for *Bacillus megaterium*
79. Jenny RJ, Mann KG, Lundblad RL (2003) A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa. Protein Expr Purif 31(1):1–11
80. Kapust RB, Waugh DS (2000) Controlled intracellular processing of fusion proteins by TEV protease. Protein Expr Purif 19(2):312–318
81. Möbius K, Arias-Cartin R, Breckau D, Hannig AL, Riedmann K, Biedendieck R, Schroder S, Becher D, Magalon A, Moser J, Jahn M, Jahn D (2010) Heme biosynthesis is coupled to electron transport chains for energy generation. Proc Natl Acad Sci U S A 107(23):10436–10441
82. Masthoff IC, David F, Wittmann C, Garnweitner G (2014) Functionalization of magnetic nanoparticles with high-binding capacity for affinity separation of therapeutic proteins. J Nanoparticle Res 16:2164
83. Martinez Cristancho CA, David F, Franco-Lara E, Seidel-Morgenstern A (2013) Discontinuous and continuous purification of single-chain antibody fragments using immobilized metal ion affinity chromatography. J Biotechnol 163(2):233–242
84. Wang W, Hollmann R, Deckwer WD (2006) Comparative proteomic analysis of high cell density cultivations with two recombinant *Bacillus megaterium* strains for the production of a heterologous dextransucrase. Proteome Sci 4:19
85. Wang W, Hollmann R, Furch T, Nimtz M, Malten M, Jahn D, Deckwer WD (2005) Proteome analysis of a recombinant *Bacillus megaterium* strain during heterologous production of a glucosyltransferase. Proteome Sci 3:4
86. Wang W, Sun J, Hollmann R, Zeng AP, Deckwer WD (2006) Proteomic characterization of transient expression and secretion of a stress-related metalloprotease in high cell density culture of *Bacillus megaterium*. J Biotechnol 126(3):313–324
87. Fürch T, Hollmann R, Wittmann C, Wang W, Deckwer WD (2007) Comparative study on central metabolic fluxes of *Bacillus megaterium* strains in continuous culture using ^{13}C labelled substrates. Bioprocess Biosyst Eng 30(1):47–59

88. Fürch T, Wittmann C, Wang W, Franco-Lara E, Jahn D, Deckwer WD (2007) Effect of different carbon sources on central metabolic fluxes and the recombinant production of a hydrolase from *Thermobifida fusca* in *Bacillus megaterium*. J Biotechnol 132(4):385–394
89. Borgmeier C, Biedendieck R, Hoffmann K, Jahn D, Meinhardt F (2011) Transcriptome profiling of *degU* expression reveals unexpected regulatory patterns in *Bacillus megaterium* and discloses new targets for optimizing expression. Appl Microbiol Biotechnol 92(3):583–596
90. Borgmeier C, Voigt B, Hecker M, Meinhardt F (2011) Functional analysis of the response regulator DegU in *Bacillus megaterium* DSM319 and comparative secretome analysis of *degSU* mutants. Appl Microbiol Biotechnol 91(3):699–711
91. Korneli C, Bolten CJ, Godard T, Franco-Lara E, Wittmann C (2012) Debottlenecking recombinant protein production in *Bacillus megaterium* under large-scale conditions-targeted precursor feeding designed from metabolomics. Biotechnol Bioeng 109(6):1538–1550
92. Wittchen KD, Meinhardt F (1995) Inactivation of the major extracellular protease from *Bacillus megaterium* DSM319 by gene replacement. Appl Microbiol Biotechnol 42(6):871–877
93. Wittchen KD, Strey J, Bultmann A, Reichenberg S, Meinhardt F (1998) Molecular characterization of the operon comprising the *spoIV* gene of *Bacillus megaterium* DSM1319 and generation of a deletion mutant. J Gen Appl Microbiol 44(5):317–326

8

Protein Complex Production in Alternative Prokaryotic Hosts

Sara Gómez, Miguel López-Estepa,
Francisco J. Fernández, and M. Cristina Vega

Abstract

Research for multiprotein expression in nonconventional bacterial and archaeal expression systems aims to exploit particular properties of “alternative” prokaryotic hosts that might make them more efficient than *E. coli* for particular applications, especially in those areas where more conventional bacterial hosts traditionally do not perform well. Currently, a wide range of products with clinical or industrial application have to be isolated from their native source, often microorganisms whose growth present numerous problems owing to very slow growth phenotypes or because they are unculturable under laboratory conditions. In those cases, transfer of the gene pathway responsible for synthesizing the product of interest into a suitable recombinant host becomes an attractive alternative solution. Despite many efforts dedicated to improving *E. coli* systems due to low cost, ease of use, and its dominant position as a ubiquitous expression host model, many alternative prokaryotic systems have been developed for heterologous protein expression mostly for biotechnological applications. Continuous research has led to improvements in expression yield through these non-conventional models, including *Pseudomonas*, *Streptomyces* and *Mycobacterium* as alternative bacterial expression hosts. Advantageous properties shared by these systems include low costs, high levels of secreted protein products and their safety of use, with non-pathogenic strains been commercialized. In addition, the use of extremophilic and halotolerant archaea as expression hosts has to be considered as a potential tool for the production of mammalian membrane proteins such as GPCRs.

S. Gómez • M. López-Estepa • F.J. Fernández
M.C. Vega (✉)
Center for Biological Research, Spanish National
Research Council (CIB-CSIC), Ramiro de Maeztu 9,
28040 Madrid, Spain
e-mail: cvega@cib.csic.es

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_8

Keywords

Alternative expression hosts • *Pseudomonas* • *Streptomyces* • *Mycobacterium* • *Archaea*

8.1 Introduction

In the socially and economically important arena of protein production the bacterium *Escherichia coli* has been immensely successful helping to overexpress tens, even hundreds, of thousands of single proteins and multi-subunit protein complexes [1]. Despite its huge success, the many advantages of *E. coli* as a protein factory are accompanied by a number of significant drawbacks or limitations, the most well-known of them being the tendency to store overexpressed but unfolded or misfolded proteins in the form of inclusion bodies. Other limitations are noticed when genes having extreme codon biases are transcribed in *E. coli*, or when specialized classes of proteins that have unusual requirements for proper folding and function are translated in the ribosome. In this chapter we will outline prokaryotic microbial factories other than *E. coli* and *Bacillus* (which are dealt with elsewhere in this book), which have been used effectively for the production of proteins, protein complexes, and complete enzymatic pathways in specific cases when the role of the expression host was crucial for success. Among a long and growing list of possible prokaryotic expression hosts we have decided to center our attention in those that are better established and/or have been used for structural biology purposes: *Pseudomonas*, *Streptomyces*, *Mycobacterium* and halophilic Archaea.

8.2 An Aerobic Multi-purpose Bacterium: *Pseudomonas*

Pseudomonas is a genus of Gram-negative aerobic soil bacteria, belonging to the family *Pseudomonadaceae*, first described by Walter Migula in 1884 and 1900 [2]. At the time of this writing, about 200 species have been identified as members of the *Pseudomonas* genus (http://www.bacterio.net/p/pseudomonas.html), with a potential broad range of applications. Based on their clinical or environmental applications, research is especially focused on the following species: *P. aeruginosa*, well known as an opportunistic human pathogen; *P. putida*, widely distributed on soil; *P. fluorescens*, a plant growth-promoting rhizobacterium with biocontrol activity against various plant pathogens; and, finally, *P. alcaligenes*, which has been used in bioremediation of oil and pesticides due to its ability to degrade polycyclic aromatic hydrocarbons. Members of the genus are characterized by their fast growth rate between a useful temperature range (30–42 °C) and an inherently high secretory capacity. In addition, many *Pseudomonas* species are oligotrophic and can thrive with very limited supply of nutrients using a multitude of different carbon sources. When these characteristics are taken into account, this bacterial group appears as an attractive source of microorganisms to develop homologous and heterologous protein expression systems, thus becoming a powerful prokaryotic cell factory and an interesting alternative to the mainstream *E. coli* systems.

8.2.1 Recombinant Protein Production in *Pseudomonas*

Given that most *Pseudomonas* spp. can be cultured in the laboratory using microbiological techniques similar to those employed with *E. coli*, most research efforts have been dedicated to identifying and characterizing promoter sequences that could be harnessed to obtain high protein yields of targeted gene products. Some of these promoters have been identified in the genomes of *Pseudomonas* spp., while other promoters have been transferred from other bacterial species, chiefly from *E. coli*, and have been shown to function across bacterial strains. A wide range of promoters has been successfully identified in *Pseudomonas* spp., both constitutive as well as inducible:

- P_L, a temperature-inducible promoter cloned from the lambda phage. Bacteria modified with a plasmid carrying the P_L promoter are regulated with a temperature-sensitive *ci857* gene. Expression control is achieved with a simple modification of culture temperature from 30 to 42 °C, which induces the denaturation of the promoter's repressor protein and, consequently, allows the expression of genes cloned downstream the P_L promoter [3].
- The strongly IPTG (isopropyl *β*-D-thiogalactoside)-inducible P*tac* promoter is widely used as a useful strategy to produce recombinant proteins in *Pseudomonas*. The P*tac* promoter is engineered as a hybrid promoter constructed from the P*trp* and P*lac* promoters from *E. coli*, and has well-known properties such as being repressed by lactose and derepressed by IPTG, a lactose analog. These properties have established the P*tac* promoter as a useful feature in expression plasmids for *E. coli*, *Pseudomonas* and other bacterial hosts, attaining high levels of expressed proteins across many different hosts [4].
- The *alk* gene cluster are involved in the metabolic degradation of alkanes is naturally present in the genome of *P. oleovorans* strains [5]. Expression vectors constructed with the *alkB* promoter, for example, can drive protein expression upon incubation with alkanes, thereby raising the interesting possibility of using such plasmids to produce alkane-degrading enzymes such as xylene oxygenase. Xylene oxygenase is involved in the production of heteroaromatic acids of interest in biotechnological processes [6].
- The regulatory control region for naphthalene and phenanthrene degradation, which is induced by sodium salicylate, has also been used as a heterologous promoter in *Pseudomonas* sp. [7].
- An induction level adjustable system for recombinant expression in *Pseudomonas* is Pm/xylS, where xylS acts as a positive regulator of Pm, which is regulated by addition of different benzoic acid derivatives to the growth medium in a range of concentrations [8].

Transformation of expression plasmids into *Pseudomonas* can be accomplished by methods that exploit all gene transfer mechanisms described in bacteria, including transformation, transduction, conjugation, and horizontal gene transfer—which is considered a central mechanism of microbial evolution [9]. The most widely employed methods to introduce foreign plasmids in *Pseudomonas* include transformation by electroporation [10] and plasmid conjugation with *E. coli* [11], with more exotic methods such as lightning transformation [12] being used to study natural electrotransformation. Owing to its simplicity of use and its high efficiency, electroporation remains as the most widespread method for transformation in *Pseudomonas*.

A broad variety of *Pseudomonas* strains are commercially available for use as hosts for protein expression. Two such strains stand out: *P. aeruginosa* PAO1-LAC and *P. putida* strain KT2440, the latter certified as generally regarded as safe (GRAS) [13]. Correct strain selection has to be considered prior to experiment in order to improve expression yields.

Vectors tailored for recombinant protein expression exist that incorporate multiple cloning sites immediately downstream to inducible promoters for regulated protein production [14]. Recently, development of shuttle vectors that can be propagated in both *Pseudomonas* and *E. coli* have appeared, constituting a strategy worth considering with which new vector libraries have been constructed (*e.g.*, UCP-Nco and pUCP-Nde). These vectors were designed with the aim to solve traditional limitations such as the need to introduce extraneous 5′ sequences to encompass translational start codons due to the impossibility to introduce NcoI or NdeI sites as direct cloning restriction sites [15].

As previously mentioned, *Pseudomonas* spp. (*e.g.*, *P. aeruginosa*) can secrete large amounts of proteins to the extracellular medium, a trait that distinguishes it from *E. coli*, where only modest levels of proteins are typically secreted. The molecular basis for this remarkable capacity lies in the presence in *P. aeruginosa*'s genome of gene loci for type II and III secretion systems [16, 17]. Unfortunately, *P. aeruginosa* is a human pathogen and molecular biology and protein expression experiments with this microorganism would require biosafety level 2 containment laboratories, which constitutes a serious limitation for routine work. To circumvent this imposing limitation, other human nonpathogenic strains that share powerful secretory machineries have been used instead. *P. fluorescens* strains are widely used to overproduce recombinant proteins due to its intrinsic ability to grow to high cellular densities while it remains less dependent on dissolved oxygen concentration. Finally, development of new genetic manipulation tools for *P. putida*, in particular for the GRAS KT2440 strain, has recently experienced a boom while researchers explore the possibilities to use it as a cell factory for biofuel production or as a source of antibody fragments [18, 19].

8.2.2 Bioremediation: Environmental Application of *Pseudomonas* Genetic Engineering

Pseudomonas has been exploited as an efficient biocontrol agent, and advances in OMICs technologies have paved the way for research to establish *Pseudomonas* as a cell factory for the production of antifungal metabolites by both natural and engineered pathways. *Pseudomonas* produces a huge range of biocontrol compounds, commonly obtained as secondary metabolites that can enhance plant health. For instance, 2,4-diacetylphloroglucinol (Phl) is a broad-spectrum antimicrobial naturally produced by *P. fluorescens* [20]. Genetic modifications of *P. fluorescens* genome have enhanced Phl production by the introduction of transcriptional regulatory control elements, and the obtained recombinant strains represent interesting alternatives to traditional chemical plant herbicides [21, 22].

Multi-protein complex expression is not extensively developed in *Pseudomonas*. However, research addressing this current limitation has made progress in areas such as bioremediation; an interesting example concerns the biodegradation of the recalcitrant compound 2-chlorotoluene. Modified strains were constructed inserting a hybrid pathway for the final biodegradation of 2-chlorotoluene to 2-chlorobenzoate, selecting as hosts two 2-chlorobenzoate degrader strains of *P. aeruginosa*. In Haro and de Lorenzo's research, the TOD [toluene dioxygenase (*tod*C1C2BA) from *P. putida* F1] and TOL pathways [encoding a benzyl alcohol dehydrogenase (*xyl*B) and a benzaldehyde dehydrogenase (*xyl*C) from pWW0 plasmid of *P. putida* mt-2] were cloned under P_u promoter regulation (inducible by toluene, xylene or analogs) in different mini-Tn5 vectors, which were separately introduced into two *P. aeruginosa* strains, JB2 and PA142. Those two 2-chlorobenzoate degrader strains express the enzymatic machinery to convert 2-chlorobenzoate into catechol. Selection of strains containing simultaneously the two plasmids was achieved growing the transformed strains on medium containing streptomycin, spectinomycin or potassium tellurite. 2-chlorotoluene bioremediation activities of the co-transformed strains were assessed by exposition to saturating 2-chlorotoluene vapors for 4 days. After this, organic compounds were extracted from the culture supernatant and analyzed by gas chromatography coupled to mass spectrometry (GC-MS). Chromatographic analysis revealed the presence of 2-chlorobenzoate, as well as many intermediates along its degradation pathway. However, no catechol final products could be unambiguously identified by this method. Although not a complete success, this innovative approach for bioremediation based on the co-transformation of several plasmids, each of which harbors complementary enzymatic activities, offers an attractive alternative route to conventional processes that should be explored further [23] (Fig. 8.1).

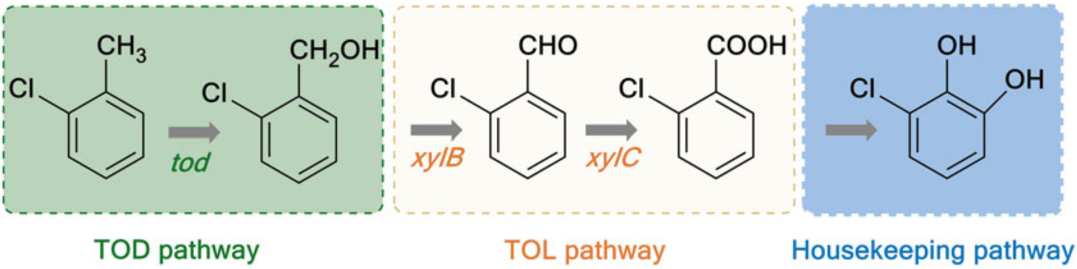


Fig. 8.1 Schematic pathway for biodegradation of 2-chlorotoluene by protein complex expression in *Pseudomonas*. Firstly, TOD pathway, encoded a toluene dioxygenase (tod), convert 2-chlorotoluene to 2-chlorobenzylalcohol. Secondly, TOL pathway, compose by two enzymes, xylB and xylC, a benzylalcohol dehydrogenase and a benzaldehyde dehydrogenase respectively, convert 2-chlorobenzylalcohol to 2-chlorobenzoate. Finally, host *P*. JB2 or PA142 provide the endogenous enzymes to consume the resulting 2-chlorobenzoate [23]

8.2.3 The Quest for New Carbon Resources

Exhaustion of existing carbon resources constitutes one of the greatest preoccupations of modern societies and chemical and energy companies, for which finding new renewable alternatives to conventional oil-based fuels is becoming a pressing need. Some of the metabolic properties of *Pseudomonas* spp., mainly the ability to grow on alkyl and aromatic organic compounds that require very specialized biochemical pathways for their assimilation, render plausible the successful development of *Pseudomonas* strains that are capable of growth under recalcitrant carbon and nitrogen sources. One possible route consists in the introduction of exogenous biochemical pathways from other bacteria into *Pseudomonas* genome. With this focus in mind, Meijnen et col. had constructed a modified *P. putida* S12 strain which carries the *xyl*XABCC operon from *Caulobacter crescentus* inserted in its genome, which endows it with a full complement of enzymes to sustain growth on D-xylose as the sole source of carbon. The endogenous activity of α-KGSA dehydrogenase is also required. *C. crescentus xyl*XABCC operon was cloned into the pJTmcs vector (selectable in medium containing gentamicin and ampicillin) under the control of the constitutive *tac* promoter. To ascertain the minimum complement of genes that is necessary to confer growth on D-xylose, a library of plasmids was obtained with the following tandem combinations of genes: XABCD (complete operon), XAD, XD, X and D. These plasmids were transformed into *P. putida* S12 and bacterial colonies grown in Luria-Bertani medium in phosphate-buffered mineral salt medium supplemented with D-glucose, D-xylose or D-xylonate as alternative carbon sources. The whole pathway was shown to be functional by spectrophotometric measurements of the D-xylose dehydrogenase and the lower pathway activity. Results suggested that only insertion of *xyl*D was sufficient to allow conversion of D-xylose to 2-ketoglutarate, whereas coexpression of *xyl*D with *xyl*XA further increased growth rates [24].

8.3 *Streptomyces*, a Factory for More Than Antibiotics

Since in 1943 Waksman and Henrici described for the first time the Actinobacteria phylum [25], *Streptomyces* spp. has undergone several important efforts to reorganize its taxonomy, mostly based on 16S ribosomal RNA sequences [26]. Despite the notorious difficulty of *Streptomyces* phylogeny and to unequivocally assign phenotypic traits to specific genes, large research programs are underway to exploit the available sequence data.

Streptomyces is a Gram-positive soil bacterial group, ubiquitous in nature, with a characteristic fungal-like growth and a high G+C content, whose members are commonly known for their role in antibiotic production. *Streptomyces* displays a unique metabolism with diverse capabilities where most compounds are produced as secondary metabolites including antibiotics, anti-cancer drugs, immunosuppressives, anti-parasitic compounds and herbicides. Their single outer membrane, combined with the intrinsic ability to secrete proteins to the supernatant in a native conformation, has propelled this genus as an attractive alternative host for the recombinant expression of proteins as well as of non-natural chemicals [27].

Currently, more than 650 species are included in this genus. Of them, *S. coelicolor* A3 is considered a useful model for genetic studies, while *S. lividans* is almost exclusively used for exogenous DNA cloning purpose [28].

8.3.1 *Streptomyces* as Host for Recombinant Protein Production

Some enzymes secreted by *Streptomyces* have industrial applications, and, thanks to advances in genetic engineering, high yielding strains as well as strains which express foreign proteins are widely developed. Advances in massive sequencing have provided whole genome sequences of *S. coelicolor* A3(2), *S. avermitilis*, *S. griseus*, *S. scabies* and *S. lividans* TK24. The knowledge gathered from genome sequences is being applied to improve recombinant protein expression in *Streptomyces* [29, 30].

Efficient protein secretion in *Streptomyces* is regulated by different translocation pathways: The Sec-dependent pathway, the twin-arginine translocation (Tat) pathway and, shared with other Actinobacteria, the ESX-1 pathway [27, 30]. Proteins targeted through each secretory pathway undergo specific signaling processing during their synthesis. In general, proteins targeted to the Sec or Tat pathways are synthesized in the form of preproteins, with a signal peptide localized at the N terminus. This signal peptide comprises a positively charged N-domain followed by a longer, hydrophobic H-domain, and a C-terminal part that contains at the end three amino acids, which form part of the signal peptidase recognition site. On the other hand, proteins targeted for secretion through the ESX-1 pathway lack any classical signal peptide, but possess a seven amino acid secretion signal at the C terminus. In all secretion pathways, during or after membrane translocation, the signal peptide is cleaved off by a specific peptidase, and the mature protein is liberated to the culture supernatant [30].

Many approaches are currently under development to increase protein secretion based on the overproduction of the components involved in the process, such as the signal peptidase [31], the Tat translocon [32], or by overexpression of PsP (phage-shock-protein) on multi-copy plasmids [33]. Modulation or modification of the signal peptide is considered among the most useful strategies to increase the yield of secreted protein product. Possible modifications include variation at the N-terminal region of the signal peptide, selection of an optimal signal peptide from a pool of alternative sequences, or even targeted optimization by directed evolution methods [34]. The latter alternative might find wider applicability since it allows the exploration of a far greater sequence space but this approach is still limited to a small group of proteins [30]. Complementary to modifying the signal peptide is the targeting of the recombinant protein through the most suitable secretory pathway. For example, selection of a Tat system seems an interesting approach when a Sec-dependent pathway does not efficiently secrete the protein of interest. Alternatively, a PsP system, which is a widely conserved mechanism of stress response, could be activated during recombinant protein production by co-expressing PsP. This method has been shown to increase the level of secreted protein.

Most vector systems designed for *Streptomyces* are based on the pIJ101 plasmid [35] that incorporates multiple constitutive promoters which are commonly used to express foreign proteins [35]. A non-exhaustive list of available constitutive promoters follows:

- *vsi*, a strong promoter from *S. venezuelae* CBS762.70, controls the expression of the highly secreted novel subtilisin inhibitor. For instance, the mouse TNFα gene was cloned under the *vsi* promoter in order to evaluate the expression and secretion mechanism of the

Table 8.1 Mammalian proteins and peptides recombinantly expressed using *S. lividans*

Strain	Protein	Size (kDa)	Yield (mg/L)	References
TK24	mTFNα (mouse)	36	200–300	[44]
TK24	IL-4R (human)	24	10	[45]
1326	α Integrin CD11b A-domain (rat)	21	8	[37]
TK24	C-terminal amidated glucagon (human)	3.5	24.2	[46]
TK24	IL-6 (human)	20	0.61	[47]

bacterial host, proving the use of *Streptomyces* as a viable expression host for mammalian proteins [36] (Table 8.1).

- *ermE-up* of *S. erythraeus* is involved in erythromycin biosynthesis. Combined use of *ermE-up* promoter and a synthetic signal peptide has allowed successful expression of a secreted and soluble recombinant rat CD11b A-domain in *S. lividans* [37].
- A metalloendopeptidase promoter isolated from *S. cinnamoneus* TH-2 was used for the construction of the plasmid pTONA5, which is based on pIJ702 vector, and used to express a secreted leucine aminopeptidase in *S. lividans* [38].
- *act1* of *S. coelicolor* CH999, from the biosynthetic gene cluster of the aromatic polyketide actinorhodin, drives protein expression to high levels when its cognate activator, ActII-ORF4, is present [39].
- The strong *kasOp** promoter of *S. coelicolor*, which controls a SARP family regulator and is strictly controlled by two regulators (ScbR and ScbR2) [40].
- *SF14p* promoter, a subcloned fragment of F14, discovered from a fragment of *S. ghanaensis* phage I19 applying the simple shotgun method with native promoters from genomic DNA.

Inducible promoters are also available for protein expression:

- *PnitA*, from the nitrilase gene of *Rhodococcus rhodochrous* J1, is strongly induced by addition of ε-caprolactam to the culture medium [41].
- Thiostrepton-inducible *PtipA* promoter of *S. lividans* [42].

Hybrid promoters are also available that were constructed to adapt the highly productive T7 expression system to *Streptomyces*. These T7-based hybrid promoters have been recently developed with the hope of harnessing the strong secretion rates of *Streptomyces* with the efficiency of the T7 RNA polymerase [43].

Several methodologies are available to introduce foreign DNA into *Streptomyces*, including transformation (with plasmid, cosmid or chromosomal DNA), transfection (with a phage replicon) or conjugation (between *Streptomyces* or *E. coli*) [35]. Chemical transformation methods based on the preparation of protoplasts used to be popular but nowadays electroporation is far more frequently used, owing to its efficiency and that it is not limited to plasmids. Unfortunately, electroporation procedures have to be optimized for mycelia from different *Streptomyces* spp. since effective electroporation conditions seem to be extremely strain-specific.

8.3.2 OMICs Technology for Improving Protein Expression

During the last two decades, huge progress in the ‘omics’ technologies has been made that might unlock the key to the production of multiple compounds. Among the ‘omics’, genomics is focused on genes, their variation and functions; transcriptomics considers information at the transcriptional level (mRNA); proteomics is centered on proteins, their expression, function and regulation; and metabolomics is focused on the metabolites produced in the organism. The application of these technologies to *Streptomyces* has so far focused in improving the production of natural antibiotics and other secondary metabolites. To this end, an enabling technology is genome shuffling, which is a method to improve the phenotype. Basically, genome shuffling consists in the construction of a recombined genome starting from multiples parental strains that are firstly subjected to several mutagenesis rounds using chemicals or physical agents. Then, protoplast fusion of mutants obtained yields multiple phenotypes and, finally, a conscientious product screening to select desired features [27, 48].

Another area of natural products research that has been aided by 'omics' technologies is the optimization of medium composition, which has a role in improving secretion yields. For example, the role played by specific amino acid supplementation was analyzed by D'huys et al. revealing an impact of amino acids on biomass growth and protein production in *S. lividans* TK24 [49]. These studies highlighted the preferred consumption by *S. lividans* TK24 of glutamate and aspartate as amino acid supplements in the expression culture, and hinted at a not well-understood correlation between high biomass and low levels of protein expression.

DNA microarrays are a useful tool to analyze in a straightforward and fast manner a wide range of genes and their patterns of expression. Antibiotics production is dependent on growth phase with the involvement of the expression of multiple genes. Genes differentially expressed (transcribed) under specified conditions are identified and manipulated to optimize strains with the aim to increase antibiotic expression level. A study done by Huang et al. has increased the knowledge about transcriptional regulation of the expression of gene clusters implicated in the biosynthesis of the antibiotics actinorhodin (Act) and undecylprodigiosin (Red), by comparing the expression profile of selected genes between different growth phases, and confirming the timely coordination between growth and antibiotic production [50].

As the cellular protein factories, ribosomes are commonly targeted for mutations aimed at modulating protein translation. Typically, resistance to antibiotics, *e.g.*, streptomycin, is one of the outcomes that are pursued through mutagenesis of the ribosomal proteins, which is especially important for *Streptomyces* strains tailored for the production of novel antibiotics, which might negatively impact on the host's translational machinery. For example, point mutations in the gene coding for the ribosomal protein rpsL in *S. lividans* can bring about an increase in protein production, resulting in an improvement of the antibiotic actinorhodin production [51].

8.3.2.1 Combinatorial Biosynthesis: An Innovative Strategy to Produce Novel Products

Nowadays there is a strong need for new compounds with antibiotic, antifungal, antiviral or antitumor activities. However, the synthesis of some of these compounds is typically hard or inaccessible, prompting researchers to find new biosynthetic procedures as well as new therapeutic molecules. Combinatorial biosynthesis is an innovative tool, based on the combination of individual metabolic reactions in order to generate a metabolic pathway capable of driving the production of novel compounds.

Using *Streptomyces* spp. as host, some combinatorial biosynthetic pathways have been constructed, for example, to synthesize indolocarbazole derivatives, which are chemical compounds with antitumor properties. Co-expression of various genes from different sources has been proposed as a viable procedure to generate many different indolocarbazole derivatives, potentially active in a tumor cell line-specific fashion. One possible pathway was assembled from genes previously isolated from rebeccamycin loci (from *Lechevalieria aerocolonigenes*), staurosporine-producing gene (*S. longisporoflavus*), as well as a halogenase gene (from *S. albogriseolus*). To perform this experiment, Sanchez et al. designed multiple *E. coli*/*Streptomyces* shuttle vectors carrying *ermE**p promoter controlling foreign gene expression, which were introduced into *S. albus* strain. Culture medium where the recombinant *S. albus* strain had grown was treated to extract and purify the compounds produced and, finally, HPLC analysis and elucidation via NMR were performed on all extracted compounds to establish a detailed biosynthetic pathway. The efficacy of the isolated compounds in eliciting antiproliferative activity was tested using a colorimetric assay against nine different cancer cell types, with promising results [52] (Fig. 8.2).

The emergence of bacterial strains resistant to a broad selection of antibiotics creates an urgent need to discover new antibiotics. Chemical/enzymatic modifications of traditional antibiotics, for example, the introduction of a sugar moiety into the antibiotic structure, should alter its biological activity positively. Deoxysugar biosynthetic gene cassettes (*des*) and a glycosyltransferase were cloned together into a replicative plasmid with thiostrepton resistance and under *ermE** regulation, obtaining a variety of gene combinations. Plasmids obtained were introduced into *S. venezuelae* YJ003, producing antibiotic analogs with unnatural sugars in their structure that were purified from culture media and analyzed by HPLC-ESI-MS and NMR. Finally, antimicrobial activity was assayed against erythromycin-resistant pathogenic strains, showing significant activities [53, 54].

In another instance, three key enzymes involved in pradimicin synthesis, a compound with antifungal and antiviral properties, were coexpressed in *S. coelicolor* CH999, and a new catalog of pradimicin analogs was generated. For this, authors developed *E. coli*/*Streptomyces* shuttle plasmids carrying PdmJ, PdmW and PdmN genes, encoding for two hydroxylases and an amino acid ligase, respectively. Results suggested than PdmJ and PdmW work collaboratively in pradimicin biosynthesis, and PdmN cannot work efficiently if cloned with only one of these enzymes, indicating that coexpression of these three enzymes is required to obtain pradimicin analogs [55].

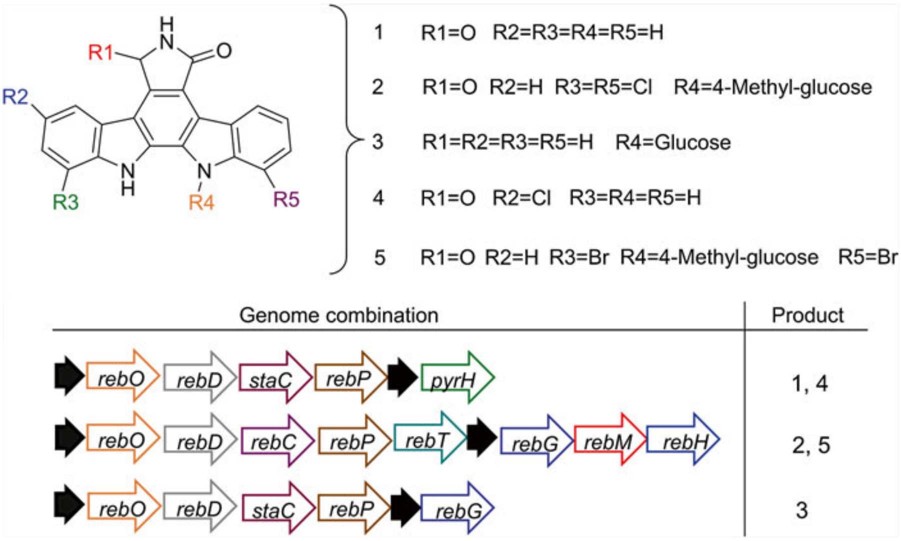

R1
H
N
O
R2
R3
N
H
N
R4
R5
1 R1=O R2=R3=R4=R5=H
2 R1=O R2=H R3=R5=Cl R4=4-Methyl-glucose
3 R1=R2=R3=R5=H R4=Glucose
4 R1=O R2=Cl R3=R4=R5=H
5 R1=O R2=H R3=Br R4=4-Methyl-glucose R5=Br
Genome combination
Product
rebO rebD staC rebP pyrH
1, 4
rebO rebD rebC rebP rebT rebG rebM rebH
2, 5
rebO rebD staC rebP rebG
3

Fig. 8.2 Multiple gene combinations assembly of the biosynthetic pathway for the synthesis of indolocarbazole derivative compounds with potentially antitumor activity. Genes involved in rebeccamycin or staurosporine biosynthesis and halogenase genes were combined to obtain a library of new indolocarbazole compounds using *S. albus* as a host. *rebO* Amino acid oxidase, *rebD* Chromopyrrolic acid synthase, *rebC* FAD-containing monooxygenase, *rebP* P450 oxygenase, *rebG* N-glycosyltransferase, *rebH* Tryptophan 7-halogenase, *rebT* Integral membrane transporter, *rebM* Sugar O-methyltransferase, *staC* FAD-containing monooxygenase, *pyrH* Tryptophan 5-halogenase [52]

8.4 Some Good Out of Bad: *Mycobacterium* as an Expression Host

As S*treptomyces*, *Mycobacteria* are a member of the Actinobacteria phylum that belongs to the family of Mycobacteriacae. They are Gram-positive, immobile bacilli, obligate aerobic, with a high genomic G+C content (59–66 %). The *Mycobacterium* genus comprises more than 170 different species, including pathogenic species that cause human diseases such as tuberculosis or leprosy. *M. tuberculosis* is responsible for around two million deaths every year, and it constitutes one of the most relevant topics of clinical investigation due to the absence of a useful treatment or vaccine and because of the emergence of drug-resistant strains [56, 57]. One limitation of these studies is the non-existence of an optimal host for high level mycobacterial protein expression. *E. coli*, widespread host for foreign protein expression, is not a successful system for mycobacterial protein expression due to some mycobacterial protein features (codon usage; restriction and modification systems; post-translational modifications in form of typical mycobacterial glycoproteins; and toxicity for product accumulation in host cells).

To solve these limitations, choosing a human non-pathogenic mycobacterial species as a source for mycobacterial pathogenic proteins is an approach to be considered. *M. bovis* (an attenuated strain) and *M. smegmatis* have been employed for the heterologous expression of mycobacterial genes as well as genes from other bacteria, viruses, and mammalian cells [58]. An extra motivation to use *M. smegmatis* to express *M. tuberculosis* proteins is that most attempts to express them in *E. coli* have led to inclusion bodies, whereas a closer bacterial species as *M. smegmatis* could in principle provide additional chaperones, factors, or an environment more conducive to the production of soluble mycobacterial proteins [59].

8.4.1 *Mycobacterium* as Host for Heterologous and Homologous Expression

Several studies have focused on identifying components of complex transcriptional regulatory systems of mycobacteria. The presence of specific promoters, transcriptional regulators and an extensive variety of sigma factors negatively affects the development of naïve protein expression strategies [60]. Those studies have identified a number of potentially useful promoters:

- *hsp60*, a strong promoter derived from *M. bovis*, is responsible for starvation stimulation, has been extensively used. Recently, serious doubts have emerged about the stability of vectors constructed with this promoter as spontaneous deletions may occur [61].
- *pBlaF* is a strong promoter that controls transcription from a β-lactamase gene from *M. fortuitum*, which has been successfully modified to improve protein export through the fusion of the gene of interest to the β-lactamase signal sequence or the inclusion of the whole β-lactamase gene as a fusion protein [62].
- Acetamidase promoter, inducible with acetamide, was isolated from *M. smegmatis*, although it is rarely used in *M. tuberculosis* due to its instability [63].
- *groEL*, encoding the mycobacterial heat shock protein GroEL2, is under the control of a promoter that allows high expression levels—under various stress conditions like nitrosative ($NaNO_2$) and heat shock (42 °C) stresses [64, 65].

For many years, research was focused on development of compatible systems with mycobacteria-*E. coli* based on shuttle vectors derived from mycobacteriophage systems such as pMV261, previously constructed based on pAL5000. For instance, pMV21, one of most widespread vector in use, comprises the kanamycin resistance gene and the *hsp60* promoter [66].

Recently, many new vectors have been elaborated. A new generation of vectors based on fosmid shuttle vectors (pMycoFos) have emerged with promising results in expression of high G+C genes and of mammalian proteins; these vectors can be propagated in *E. coli* with a controllable copy number and expression of target genes can be induced in *Mycobacterium* spp. [63]. Another approach with higher yields is the possibility to re-engineer a shuttle vector modifying the promoter with the strong lacZ sequence [67]. Development of inducible vectors, controlled by addition of substrates to the culture medium [68] or of temperature-sensitive mutant strains, promises to generate new vectors that could improve the mycobacterium expression possibilities. For instance, using mutagenesis of pAL5000 a thermosensitive plasmid was obtained that could be stably maintained if culture was grown at 32 °C or below [69].

Mycobacterium spp. present the disadvantage of being difficult to transform. Generally, electroporation is the most successful method to introduce plasmid DNA into mycobacteria. In addition, plasmids can also be introduced by conjugation [70] or through mycobacteriophages [71]. A huge variety of parameters may influence transformation efficiency, including, among others, strain, selection marker used and conditions for the electroporation pulse. Some protocols are optimized for commonly used strains such as *M. tuberculosis* or *M. smegmatis* [58].

An interesting though little exploited aspect of mycobacteria is their capacity, shared with many Gram-positive bacteria, to secrete proteins directly into the culture medium through a specialized type VII secretion system (also called ESX secretion pathway). The mycobacterial ESX secretion pathway remains poorly understood. Four ESX loci have been described that encode for PE and PPE proteins, which bear a conserved secretion signal at the C terminus that is absolutely required for secretion [72]. Further developments in our understanding of the mycobacterial-specific ESX system will certainly facilitate the exploitation of the secretion pathway for recombinant protein production.

8.4.2 Remarks on Complex Protein Expression

As far as we know no example of a co-expression experiment has been published to date using a mycobacterial expression host that demonstrates expression of large amounts of a protein complex. However, recent developments in vaccine production suggest that co-expression plasmids could be a useful addition to the available collection of strains and vectors. For example, shuttle *E. coli*/*M. tuberculosis* vectors have been constructed to co-express antigen 85B from *M. tuberculosis* and regulatory trans-activating protein from the HIV virus in order to develop new generation vaccines for co-immunizing against both diseases. The expression host selected was *E. coli*, although using *M. smegmatis* instead could represent a breakthrough and place it as a useful cell factory for vaccine development [73].

Recently, Parikh et col. have developed a catalog of shuttle vector for the constitutive expression of two proteins independently controlled by two promoters. Plasmids were constructed with two multiple cloning sites under P*smyc* (without *tet* operator) or P*myc-tet*O promoters (with *tet* operator, that acts as constitutive in the absence of the repressor). To facilitate further purification, proteins expressed under P*smyc* promoter were fused to a His-tag, while proteins encoded under P*myc-tet*O promoter were fused to a FLAG-tag. In that case, two groups of proteins were tested for coexpression: a pair formed by *pkn*B (kinase) and its cognate substrate *gar*A, and a pair formed by *pkn*K and its substrate *vir*S. All genes were amplified from *M. tuberculosis* H37Rv bacterial artificial chromosome (BAC) clones. Introduced in *M. smegmatis* by electroporation in both cases, expression was analyzed by Western blot, confirming that the vectors carried all the elements for coexpression, including kanamycin as selectable marker for positive clones [74].

Designing an expression system that allows differential control of multicomponent complexes is the goal of the work carried out by Chang et al. With a focus on obtaining a vector whose expression levels could be tunable, the authors described how to construct a single vector for co-expression of two genes for protein-protein interaction analysis in mycobacteria. In particular, *M. smegmatis* was transformed with a vector carrying a hygromycin resistance gene and two genes of the stearoyl-CoAΔ9 desaturase complex: rv3229c gene (encoding stearoyl-CoA desaturase, DesA3) and rv3230c gene (encoding NADPH:stearoyl-CoA desaturase oxidoreductase, Rv3230c) under the control of P*hsp* constitutive promoter and P*tetO* inducible promoter, respectively. Expression of the genes under P*tetO* in plasmid pTetCoex was enhanced by the addition of an unrelated gene located 3′ to the gene of interest, which increased mRNA stability and avoided mRNA degradation. Cells transformed with this vector were able to co-express both proteins, which was confirmed with an enzymatic assay. The main challenge that this work overcame was the previously observed toxicity caused by expression of Rv3230, which was solved using an inducible system; in contrast, constitutive accumulation of non-toxic DesA3 was observed. Upon induction, the expression levels of Rv3230 were increased, thereby allowing the formation of functional complex [75].

8.5 Worth Their Salt: Archaeal Expression Systems

Archaea, sometimes referred to as the third branch of life, is a special kingdom of single-celled microorganisms that has properties that separates them from bacterial and eukaryotic organisms. The Archaea are divided into four phyla and, although much of earlier work focused on extremophilic archaeal organisms isolated from hot springs and salt lakes, they are recognized today to populate a broad range of habitats spanning soils, oceans, marshlands and the human colon and navel.

Within the Archaea, the family *Halobacteriaceae* belongs to the *Euryarcheota* phylum, an extremely halophilic group, which includes aerobic or facultative anaerobic organisms, generally red-pigmented. As the halophiles *par excellence*, their optimal growing conditions include hypersaline media although, in contrast to other Archaea, *Halobacteria* have the enormous advantage of being able to grow in standard laboratory media supplemented with salt [76]. Many strains have been fully sequenced, showing a relatively high G+C content (58–68 %) and the presence of large and small extrachromosomal elements. The combination of easy culturing, knowledge of their interesting biology, and development of tools for genome manipulation has transformed this group into a fascinating model organism [77]. However, some peculiarities have to be taken into account with *Halobacteria* that have implications for their use in experimental work: Their genomes tend to be genetically unstable due to a very high number of active insertion sequences and transposons; and show slow growth rates and a great sensitivity to lysis during transformation.

8.5.1 Archaeal Genetics Methodology

Some halobacterial species are regarded as excellent models to understand biological questions that transcend the archaeal domain. This is well exemplified by their extensive use to study the structure and function of membrane transport systems, based on the fact that many halobacterial species naturally produce large amounts of transmembrane proteins termed bacteriorhodopsins (BR), which are excellent model proteins for mammalian GPCRs. Nevertheless, although some genetic tools (transformation protocols, vectors, selectable markers, etc.) are available, further research will be necessary to widen the applicability of Archaea as model systems [78].

8.5.1.1 Transformation

The most common method to introduce foreign DNA plasmids into Archaea is based on polyethylene glycol (PEG)-based chemical transformation of spheroplasts, which can be previously prepared through divalent cation chelation with EDTA, followed by cell regeneration after DNA uptake [79, 80].

8.5.1.2 Selectable Markers and Reporter Genes

The most common negative selection marker in Archaea is the *hmg*/*mev* gene, which confers resistance to the antibiotic mevinolin. Mevinolin is an inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A reductase, an enzyme involved in the biosynthesis of isoprenoid chains that form an essential part of membrane lipids in Archaea. Other commonly used selection schemes include resistance to 5-fluoroorotic acid (resistance gene *ura3*/*purF* and *ura5*/*pyrE*) and to novobiocin (encoded by the *gyrB* resistance gene) [78]. In addition to available antibiotic resistance markers, several reporter genes are also available, among which the following are the most frequently used: bacterio-opsin gene (*bop*), a halophilic β-galactosidase gene (*bga*H), and GFP [80].

8.5.1.3 Shuttle and Cloning Vectors

Many archaeal vectors are developed based on the huge plasmid diversity intrinsically present in this group. Plasmids pNRC100 and pHH1 are useful vector derivatives from *Halobacterium* spp. that have the advantage of being capable of replicating in all strains of halobacteria. Shuttle *E. coli*-*Halobacterium* plasmids can be constructed using appropriate resistance markers, *e.g.*, mevinolin resistance in haloarchaea and ampicillin/kanamycin resistance in *E. coli* [78].

8.5.1.4 Promoters

Suitable promoters are selected firstly based on the target protein. When the protein to be expressed is expected to be well folded and soluble, a strong, constitutive promoter such as ferredoxin (*fdx*) promoter might be the first option, and the gene of interest can be fused to a His-tag to allow affinity purification of the expressed protein. In contrast, for the production of membrane proteins in *Halobacterium*, the inducible *bop* (bacterio-opsin) promoter has yielded the best results. The *bop* promoter is induced under low-oxygen tension and by high light intensity [81]. For example, expression of mammalian transmembrane GPCRs, such as the muscarinic acetylcholine and the adrenergic receptors, was carried out using a *bop* fusion strategy under *bop* promoter in *Haloferax volcanii*, with functional receptor expression detected. The *bop* fusion strategy exploits the high expression levels associated with bacteriorhodopsin by the fusion of the target cDNA with regulatory and translational *bop* sequences such as promoter, transcriptional terminator or regulatory factor binding sites [82].

Other alternative systems have been tried. For example, the gas vesicle gene cluster (*gvpC*) was modified to produce fusion proteins directly attached to the surface of floating vesicles in *Halobacterium salinarum*. Proteins expressed under this promoter showed lower yields than with other available promoters, but this might be due to the limited knowledge currently available about the regulation of the *gvpC* promoter [83]. More recently, a new inducible system has been developed with promising results, based on the use of the *kpd* promoter, which is inducible by K^+, and using *bgaH* (β-galactosidase from Archaea) as a reporter for testing gene regulation [84].

8.5.2 Complex Protein Expression in Archaea

A very interesting case of co-expression of recombinant proteins in an archaeal host was recently published [85]. In this work, the authors expressed multiple proteins that were able to interact with the intrinsic host-encoded machinery to restore phototaxis. In this case, co-expression of the recombinant protein from *Natronobacterium pharaonis* into *H. salinarum* produces a motility response via the flagella of the host. Authors could show overproduction of psRII (encoding the phototactic receptor sensory rhodopsin II) and HtrII (a halobacterial transducer of sensory rhodopsins), both cloned from the high halotolerant *N. pharaonis*, into the expression system *H. salinarum* Pho81/w, a carotenoid deficient, knock-out for the rhodopsins and transducers HtrI and II and non-sensitive to the light [85–88]. The genes encoding both proteins, phtrII (HtrII) and psopII (psRII), were placed in the same plasmid (pNphtrII/psopII, an

H. salinarum/*E. coli* shuttle vector) under the control of the *bop* promoter, which was then induced by exposure to light or under oxygen-limitation, and carrying a novobiocin resistance gene as selectable marker [89, 90]. This construct was inserted in the *H. salinarum* genome by homologous recombination via the *bop* locus. Positive colonies, checked by PCR, were then analyzed by determining the expression level of recombinant protein psRII by a spectroscopy method [85]. To confirm co-expression, a motility experiment was performed that showed that only cells successfully co-transformed were able to move away from the light, thereby restoring negative phototaxis from wild type (Fig. 8.3).

Another area of biotechnology where simultaneous expression of multiple proteins has been assayed in an archaeal host is in bioplastic production. The need to find sustainable alternatives to oil derivatives, such as the biodegradable PHA (polyesters of hydroxyalkanoates), has drawn much attention in recent years and efforts to develop new methods to produce them are of great importance [91]. Some authors have described the bioproduction of PHA using recombinant haloarchaea, based on the original report in 1972 that this microorganism could produce such compounds naturally [92, 93].

However, these compounds have remained very expensive to produce, although recent promising applications of this material to manufacture medical devices has brought this topic to the forefront [94]. An enzyme implicated in the production of PHA is the PHA synthase, which requires coenzyme A thioesters of hydroxyalkanoic acids to synthesize the final product of interest [95]. It is known that *Allochromatium vinosum* (*Chromatium vinosum*), an anaerobic photosynthetic purple sulfur bacterium, has the ability to produce PHB (polyhidroxybutyrate), another kind of PHA, which is carried out by a PHA synthase formed by two subunits (heterodimer), PhaE and PhaC [96]. Although some Archaea have shown the same ability in PHB production, no additional research was carried out in this until Han et al. [92] characterized the homologous sequence of *phaEC$_H$m* operon from *H. marismortui*, which carries the genetic informa-

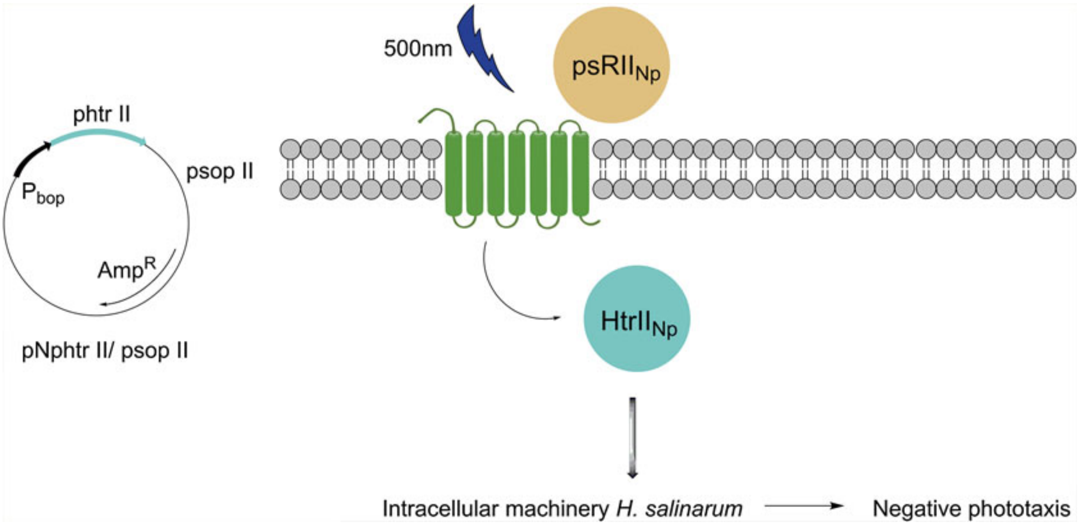
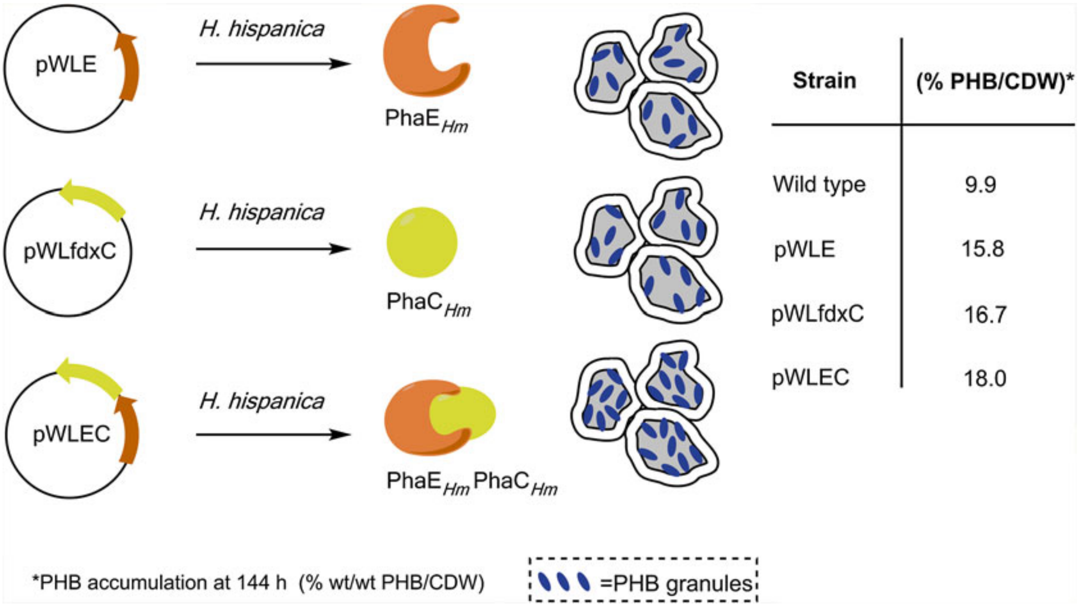


Fig. 8.3 Coexpression of proteins restored phototaxis in Archaea. Illustration of the interaction between the recombinant proteins from *N. pharaonis* and the intracellular machinery of the host (*H. salinarum*) to transform the external signal (light at 500 nm) in a mechanical response (negative phototaxis). The plasmid phtrII/psopII, harboring the psop II and phtr II genes and encoding psRII$_{NP}$ and HtrII$_{NP}$ proteins respectively, was transformed into *H. salinarium*, in order to restore truncated phototaxis genes. *psRII$_{NP}$* (psopII) phototactic receptors sensory rhodopsin II, *HtrII$_{NP}$* (phtrII) halobacterial transducers of sensory rhodopsins, both from *N. pharaonis*, *P$_{bop}$* bacterio-opsin promoter, *Amp^R* ampillicin resistance [85]



Strain	(% PHB/CDW)*
Wild type	9.9
pWLE	15.8
pWLfdxC	16.7
pWLEC	18.0

Fig. 8.4 Biosynthesis of PHB by coexpression in Archaea. Illustration of the overexpression of the proteins PhaE and PhaC from *H. marismortui* using *H. hispanica* ATCC33960 as an expression system. Coexpression of PhaE and PhaC results in higher amounts of PHB accumulation (showed as *blue* granulates). PHA enzymes (*$PhaE_{HM}$ and $PhaC_{HM}$*) polyesters of hydroxyalkanoates synthases, *PHB* polyhidroxybutyrate, *CDW* cellular dry weight. *: PHB accumulation calculated as % wt/wt between PHB/CDW measured at 144 h [92]

tion for the synthesis of PHB. The authors designed three constructs based on shuttle vectors to clone these genes (isolated or in conjunction) into *H. hispanica*, a species phylogenetically close to *H. marismortui* but much easier to transform, and carrier of the *$phaEC_{Hh}$* operon which has high homology with the *$phaEC_{H}m$* operon of *H. marismortui*. The following expression plasmids were constructed based on the original shuttle vector pWL102, which carries ampicillin and mevinolin resistance genes: pWLE, encoding PhaE_{Hm} under its native *pha* promoter; pWLEC, encoding PhaE_{Hm} and PhaC_{Hm} also under *pha* promoter; and pWLfdxC, encoded PhaC_{Hm} under *fdxHm* promoter. Transformation with the co-expression plasmids into *H. hispanica phaEC* knock-out cell strain was done using a PEG (polyethylene glycol)-based protocol [97]. Results of the co-expression experiments suggested that only cells transformed with pWLEC, encoding both proteins, were able to overproduce and accumulate PHB, showing that both proteins are implicated directly and act synergistically in the synthesis of the PHB [92] (Fig. 8.4).

8.6 Conclusions

The success of a protein expression experiment frequently depends critically on the choice of expression host. Many useful prokaryotic hosts besides *E. coli* are readily available for use and some of them might even perform better than *E. coli* for certain protein classes or when the desired outcome is a small molecule or an antibiotic. Expression systems based on *Pseudomonas*, *Streptomyces*, *Mycobacterium* and *Halobacterium*, among others, have been shown to provide alternative hosts for the production of a rich variety of prokaryotic and mammalian proteins as well as factories for fine chemicals. Undoubtedly, future research in genomics and proteomics of non-conventional model microorganisms will increase their roles in the biotechnology and pharmaceutical production of proteins and protein complexes.

Acknowledgments The authors received funding from the Spanish Ministry of Science and Innovation grant CTQ2015-66206-C2-2-R and the European Community's Seventh Framework Programme (FP7/2007–2013) for the ComplexINC Project under grant agreement N° 279039.

References

1. Fernandez FJ, Vega MC (2013) Technologies to keep an eye on: alternative hosts for protein production in structural biology. Curr Opin Struct Biol 23(3):365–373
2. Migula W (1900) System der Bakterien, vol 2. Gustav Fischer, Jena
3. Milman G (1987) Expression plasmid containing the lambda PL promoter and cI857 repressor. Methods Enzymol 153:482–491
4. Liu Y, Rainey PB, Zhang X-X (2014) Mini-Tn7 vectors for studying post-transcriptional gene expression in Pseudomonas. J Microbiol Methods 107:182–185
5. van Beilen JB, Panke S, Lucchini S, Franchini AG, Rothlisberger M, Witholt B (2001) Analysis of Pseudomonas putida alkane-degradation gene clusters and flanking insertion sequences: evolution and regulation of the alk genes. Microbiology 147(Pt 6):1621–1630
6. Panke S, Meyer A, Huber CM, Witholt B, Wubbolts MG (1999) An alkane-responsive expression system for the production of fine chemicals. Appl Environ Microbiol 65(6):2324–2332
7. Husken LE, Beeftink R, de Bont JA, Wery J (2001) High-rate 3-methylcatechol production in Pseudomonas putida strains by means of a novel expression system. Appl Microbiol Biotechnol 55(5):571–577
8. Steigedal M, Valla S (2008) The Acinetobacter sp. chnB promoter together with its cognate positive regulator ChnR is an attractive new candidate for metabolic engineering applications in bacteria. Metab Eng 10(2):121–129
9. Davison J (1999) Genetic exchange between bacteria in the environment. Plasmid 42(2):73–91
10. Filiatrault MJ, Stodghill PV, Wilson J, Butcher BG, Chen H, Myers CR, Cartinhour SW (2013) CrcZ and CrcX regulate carbon source utilization in Pseudomonas syringae pathovar tomato strain DC3000. RNA Biol 10(2):245–255
11. De Laurentis W, Leang K, Hahn K, Podemski B, Adam A, Kroschwald S, Carter LG, van Pee K-H, Naismith JH (2006) Preliminary crystallographic characterization of PrnB, the second enzyme in the pyrrolnitrin biosynthetic pathway. Acta Crystallogr Sect F: Struct Biol Cryst Commun 62(Pt 11):1134–1137
12. Ceremonie H, Buret F, Simonet P, Vogel TM (2006) Natural Pseudomonas sp. strain N3 in artificial soil microcosms. Appl Environ Microbiol 72(4):2385–2389
13. Martinez-Garcia E, Nikel PI, Aparicio T, de Lorenzo V (2014) Pseudomonas 2.0: genetic upgrading of P. putida KT2440 as an enhanced host for heterologous gene expression. Microb Cell Fact 13(1):159–159
14. West SE, Schweizer HP, Dall C, Sample AK, Runyen-Janecky LJ (1994) Construction of improved Escherichia-Pseudomonas shuttle vectors derived from pUC18/19 and sequence of the region required for their replication in Pseudomonas aeruginosa. Gene 148(1):81–86
15. Cronin CN, McIntire WS (1999) PUCP-Nco and pUCP-Nde: Escherichia-Pseudomonas shuttle vectors for recombinant protein expression in Pseudomonas. Anal Biochem 272(1):112–115
16. Derouazi M, Toussaint B, Quenee L, Epaulard O, Guillaume M, Marlu R, Polack B (2008) High-yield production of secreted active proteins by the Pseudomonas aeruginosa type III secretion system. Appl Environ Microbiol 74(11):3601–3604
17. Krzeslak J, Braun P, Voulhoux R, Cool RH, Quax WJ (2009) Heterologous production of Escherichia coli penicillin G acylase in Pseudomonas aeruginosa. J Biotechnol 142(3–4):250–258
18. Nikel PI, de Lorenzo V (2014) Robustness of Pseudomonas putida KT2440 as a host for ethanol biosynthesis. N Biotechnol 31(6):562–571
19. Dammeyer T, Steinwand M, Kruger S-C, Dubel S, Hust M, Timmis KN (2011) Efficient production of soluble recombinant single chain Fv fragments by a Pseudomonas putida strain KT2440 cell factory. Microb Cell Fact 10:11–11
20. Mark G, Morrissey JP, Higgins P, O'Gara F (2006) Molecular-based strategies to exploit Pseudomonas biocontrol strains for environmental biotechnology applications. FEMS Microbiol Ecol 56(2):167–177
21. Hofte M, Altier N (2010) Fluorescent pseudomonads as biocontrol agents for sustainable agricultural systems. Res Microbiol 161(6):464–471
22. Weller DM, Landa BB, Mavrodi OV, Schroeder KL, De La Fuente L, Blouin Bankhead S, Allende Molar R, Bonsall RF, Mavrodi DV, Thomashow LS (2007) Role of 2,4-diacetylphloroglucinol-producing fluorescent Pseudomonas spp. in the defense of plant roots. Plant Biol 9(1):4–20
23. Haro M-A, de Lorenzo V (2001) Metabolic engineering of bacteria for environmental applications: construction of Pseudomonas strains for biodegradation of 2-chlorotoluene. J Biotechnol 85(2):103–113
24. Meijnen JP, de Winde JH, Ruijssenaars HJ (2009) Establishment of oxidative D-xylose metabolism in Pseudomonas putida S12. Appl Environ Microbiol 75(9):2784–2791
25. Waksman SA, Henrici AT (1943) The nomenclature and classification of the actinomycetes. J Bacteriol 46(4):337–341

26. Wellington EM, Stackebrandt E, Sanders D, Wolstrup J, Jorgensen NO (1992) Taxonomic status of Kitasatosporia, and proposed unification with Streptomyces on the basis of phenotypic and 16S rRNA analysis and emendation of Streptomyces Waksman and Henrici 1943, 339AL. Int J Syst Bacteriol 42(1):156–160
27. Anne J, Maldonado B, Van Impe J, Van Mellaert L, Bernaerts K (2012) Recombinant protein production and streptomycetes. J Biotechnol 158(4):159–167
28. Anne J, Van Mellaert L (1993) Streptomyces lividans as host for heterologous protein production. FEMS Microbiol Lett 114(2):121–128
29. Ruckert C, Albersmeier A, Busche T, Jaenicke S, Winkler A, Friethjonsson OH, Hreggviethsson GO, Lambert C, Badcock D, Bernaerts K, Anne J, Economou A, Kalinowski J (2015) Complete genome sequence of Streptomyces lividans TK24. J Biotechnol 199:21–22
30. Anne J, Vrancken K, Van Mellaert L, Van Impe J, Bernaerts K (2014) Protein secretion biotechnology in Gram-positive bacteria with special emphasis on Streptomyces lividans. Biochim Biophys Acta 1843(8):1750–1761
31. Palacin A, Parro V, Geukens N, Anne J, Mellado RP (2002) SipY Is the Streptomyces lividans type I signal peptidase exerting a major effect on protein secretion. J Bacteriol 184(17):4875–4880
32. Barrett CM, Ray N, Thomas JD, Robinson C, Bolhuis A (2003) Quantitative export of a reporter protein, GFP, by the twin-arginine translocation pathway in Escherichia coli. Biochem Biophys Res Commun 304(2):279–284
33. Wang YY, Fu ZB, Ng KL, Lam CC, Chan AK, Sze KF, Wong WK (2011) Enhancement of excretory production of an exoglucanase from Escherichia coli with phage shock protein A (PspA) overexpression. J Microbiol Biotechnol 21(6):637–645
34. Lammertyn E, Anne J (1998) Modifications of Streptomyces signal peptides and their effects on protein production and secretion. FEMS Microbiol Lett 160(1):1–10
35. Kieser TBM, Buttner MJ, Chater KF, Hopwood DA (2000) Practical Streptomyces genetics. The John Innes Foundation, Norwich
36. Lammertyn E, Van Mellaert L, Schacht S, Dillen C, Sablon E, Van Broekhoven A, Anne J (1997) Evaluation of a novel subtilisin inhibitor gene and mutant derivatives for the expression and secretion of mouse tumor necrosis factor alpha by Streptomyces lividans. Appl Environ Microbiol 63(5):1808–1813
37. Ayadi DZ, Chouayekh H, Mhiri S, Zerria K, Fathallah DM, Bejar S (2007) Expression by streptomyces lividans of the rat alpha integrin CD11b A-domain as a secreted and soluble recombinant protein. J Biomed Biotechnol 2007(1):54327
38. Hatanaka T, Onaka H, Arima J, Uraji M, Uesugi Y, Usuki H, Nishimoto Y, Iwabuchi M (2008) pTONA5: a hyperexpression vector in Streptomycetes. Protein Expr Purif 62(2):244–248
39. Rowe CJ, Cortés J, Gaisser S, Staunton J, Leadlay PF (1998) Construction of new vectors for high-level expression in actinomycetes. Gene 216(1):215–223
40. Wang W, Li X, Wang J, Xiang S, Feng X, Yang K (2013) An engineered strong promoter for streptomycetes. Appl Environ Microbiol 79(14):4484–4492
41. Herai S, Hashimoto Y, Higashibata H, Maseda H, Ikeda H, Omura S, Kobayashi M (2004) Hyper-inducible expression system for streptomycetes. Proc Natl Acad Sci U S A 101(39):14031–14035
42. Murakami T, Holt TG, Thompson CJ (1989) Thiostrepton-induced gene expression in Streptomyces lividans. J Bacteriol 171(3):1459–1466
43. Lussier FX, Denis F, Shareck F (2010) Adaptation of the highly productive T7 expression system to Streptomyces lividans. Appl Environ Microbiol 76(3):967–970
44. Pozidis C, Lammertyn E, Politou AS, Anne J, Tsiftsoglou AS, Sianidis G, Economou A (2001) Protein secretion biotechnology using Streptomyces lividans: large-scale production of functional trimeric tumor necrosis factor alpha. Biotechnol Bioeng 72(6):611–619
45. Zhang Y, Wang WC, Li Y (2004) Cloning, expression, and purification of soluble human interleukin-4 receptor in Streptomyces. Protein Expr Purif 36(1):139–145
46. Qi X, Jiang R, Yao C, Zhang R, Li Y (2008) Expression, purification, and characterization of C-terminal amidated glucagon in Streptomyces lividans. J Microbiol Biotechnol 18(6):1076–1080
47. Zhu Y, Wang L, Du Y, Wang S, Yu T, Hong B (2011) Heterologous expression of human interleukin-6 in Streptomyces lividans TK24 using novel secretory expression vectors. Biotechnol Lett 33(2):253–261
48. Chaudhary AK, Dhakal D, Sohng JK (2013) An insight into the "-omics" based engineering of streptomycetes for secondary metabolite overproduction. Biomed Res Int 2013:968518
49. D'Huys PJ, Lule I, Van Hove S, Vercammen D, Wouters C, Bernaerts K, Anne J, Van Impe JF (2011) Amino acid uptake profiling of wild type and recombinant Streptomyces lividans TK24 batch fermentations. J Biotechnol 152(4):132–143
50. Huang J, Lih C-J, Pan K-H, Cohen SN (2001) Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in Streptomyces coelicolor using DNA microarrays. Genes Dev 15(23):3183–3192
51. Okamoto S, Lezhava A, Hosaka T, Okamoto-Hosoya Y, Ochi K (2003) Enhanced expression of S-adenosylmethionine synthetase causes overproduction of actinorhodin in Streptomyces coelicolor A3(2). J Bacteriol 185(2):601–609
52. Sanchez C, Zhu L, Brana AF, Salas AP, Rohr J, Mendez C, Salas JA (2005) Combinatorial biosynthe-

sis of antitumor indolocarbazole compounds. Proc Natl Acad Sci U S A 102(2):461–466

53. Salas JA, Mendez C (2007) Engineering the glycosylation of natural products in actinomycetes. Trends Microbiol 15(5):219–232
54. Shinde PB, Han AR, Cho J, Lee SR, Ban YH, Yoo YJ, Kim EJ, Kim E, Song MC, Park JW, Lee DG, Yoon YJ (2013) Combinatorial biosynthesis and antibacterial evaluation of glycosylated derivatives of 12-membered macrolide antibiotic YC-17. J Biotechnol 168(2):142–148
55. Napan K, Zhang S, Morgan W, Anderson T, Takemoto JY, Zhan J (2014) Synergistic actions of tailoring enzymes in pradimicin biosynthesis. Chembiochem Eur J Chem Biol 15(15):2289–2296
56. Vitoria M, Granich R, Gilks CF, Gunneberg C, Hosseini M, Were W, Raviglione M, De Cock KM (2009) The global fight against HIV/AIDS, tuberculosis, and malaria: current status and future perspectives. Am J Clin Pathol 131(6):844–848
57. Wright A, Zignol M, Van Deun A, Falzon D, Gerdes SR, Feldman K, Hoffner S, Drobniewski F, Barrera L, van Soolingen D, Boulabhal F, Paramasivan CN, Kam KM, Mitarai S, Nunn P, Raviglione M, Global Project on Anti-Tuberculosis Drug Resistance S (2009) Epidemiology of antituberculosis drug resistance 2002–07: an updated analysis of the Global Project on Anti-Tuberculosis Drug Resistance Surveillance. Lancet 373(9678):1861–1873
58. Parish T, Brown AC (2009) Mycobacteria protocols. Methods in Molecular Biology, vol 465. Humana Press, New York
59. Goldstone RM, Moreland NJ, Bashiri G, Baker EN, Shaun Lott J (2008) A new Gateway vector and expression protocol for fast and efficient recombinant protein expression in Mycobacterium smegmatis. Protein Expr Purif 57(1):81–87
60. Newton-Foot M, Gey van Pittius NC (2013) The complex architecture of mycobacterial promoters. Tuberculosis 93(1):60–74
61. Al-Zarouni M, Dale JW (2002) Expression of foreign genes in Mycobacterium bovis BCG strains using different promoters reveals instability of the hsp60 promoter for expression of foreign genes in Mycobacterium bovis BCG strains. Tuberculosis 82(6):283–291
62. Nascimento IP, Dias WO, Mazzantini RP, Miyaji EN, Gamberini M, Quintilio W, Gebara VC, Cardoso DF, Ho PL, Raw I, Winter N, Gicquel B, Rappuoli R, Leite LC (2000) Recombinant Mycobacterium bovis BCG expressing pertussis toxin subunit S1 induces protection against an intracerebral challenge with live Bordetella pertussis in mice. Infect Immun 68(9):4877–4883
63. Ly MA, Liew EF, Le NB, Coleman NV (2011) Construction and evaluation of pMycoFos, a fosmid shuttle vector for Mycobacterium spp. with inducible gene expression and copy number control. J Microbiol Methods 86(3):320–326
64. Stover CK, de la Cruz VF, Fuerst TR, Burlein JE, Benson LA, Bennett LT, Bansal GP, Young JF, Lee MH, Hatfull GF et al (1991) New use of BCG for recombinant vaccines. Nature 351(6326):456–460
65. Joseph SV, Madhavilatha GK, Kumar RA, Mundayoor S (2012) Comparative analysis of mycobacterial truncated hemoglobin promoters and the groEL2 promoter in free-living and intracellular mycobacteria. Appl Environ Microbiol 78(18):6499–6506
66. Hatfull GF (2014) Molecular genetics of mycobacteriophages. Microbiol Spectr 2(2):1–36
67. Eitson JL, Medeiros JJ, Hoover AR, Srivastava S, Roybal KT, Ainsa JA, Hansen EJ, Gumbo T, van Oers NS (2012) Mycobacterial shuttle vectors designed for high-level protein expression in infected macrophages. Appl Environ Microbiol 78(19):6829–6837
68. Williams KJ, Joyce G, Robertson BD (2010) Improved mycobacterial tetracycline inducible vectors. Plasmid 64(2):69–73
69. Guilhot C, Gicquel B, Martin C (1992) Temperature-sensitive mutants of the Mycobacterium plasmid pAL5000. FEMS Microbiol Lett 77(1–3):181–186
70. Gormley EP, Davies J (1991) Transfer of plasmid RSF1010 by conjugation from Escherichia coli to Streptomyces lividans and Mycobacterium smegmatis. J Bacteriol 173(21):6705–6708
71. Bardarov S, Kriakov J, Carriere C, Yu S, Vaamonde C, McAdam RA, Bloom BR, Hatfull GF, Jacobs WR Jr (1997) Conditionally replicating mycobacteriophages: a system for transposon delivery to Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 94(20):10961–10966
72. Daleke MH, Ummels R, Bawono P, Heringa J, Vandenbroucke-Grauls CM, Luirink J, Bitter W (2012) General secretion signal for the mycobacterial type VII secretion pathway. Proc Natl Acad Sci U S A 109(28):11342–11347
73. Pardini M, Giannoni F, Palma C, Iona E, Cafaro A, Brunori L, Rinaldi M, Fazio VM, Laguardia ME, Carbonella DC, Magnani M, Ensoli B, Fattorini L, Cassone A (2006) Immune response and protection by DNA vaccines expressing antigen 85B of Mycobacterium tuberculosis. FEMS Microbiol Lett 262(2):210–215
74. Parikh A, Kumar D, Chawla Y, Kurthkoti K, Khan S, Varshney U, Nandicoori VK (2013) Development of a new generation of vectors for gene expression, gene replacement, and protein-protein interaction studies in mycobacteria. Appl Environ Microbiol 79(5):1718–1729
75. Chang Y, Mead D, Dhodda V, Brumm P, Fox BG (2009) One-plasmid tunable coexpression for mycobacterial protein-protein interaction studies. Protein Sci Publ Protein Soc 18(11):2316–2325
76. Oren A (2014) Taxonomy of halophilic Archaea: current status and future challenges. Extremophiles 18(5):825–834

77. Soppa J (2006) From genomes to function: haloarchaea as model organisms. Microbiology 152(Pt 3):585–590
78. Berquist BR, Müller JA, DasSarma S (2006) Genetic systems for halophilic archaea, vol 35, Methods in microbiology. Academic Press, Cambridge, MA
79. Cline SW, Schalkwyk LC, Doolittle WF (1989) Transformation of the archaebacterium Halobacterium volcanii with genomic DNA. J Bacteriol 171(9):4987–4991
80. Dyall-Smith M (2009) The halohandbook: protocols for haloarchaeal genetics (version 7.2). Available online at: http://www.haloarchaea.com/resources/halohandbook/
81. Gregor D, Pfeifer F (2005) In vivo analyses of constitutive and regulated promoters in halophilic archaea. Microbiology 151(Pt 1):25–33
82. Bartus CL, Jaakola V-P, Reusch R, Valentine HH, Heikinheimo P, Levay A, Potter LT, Heimo H, Goldman A, Turner GJ (2003) Downstream coding region determinants of bacterio-opsin, muscarinic acetylcholine receptor and adrenergic receptor expression in Halobacterium salinarum. BBA Biomembr 1610(1):109–123
83. Bleiholder A, Frommherz R, Teufel K, Pfeifer F (2012) Expression of multiple tfb genes in different Halobacterium salinarum strains and interaction of TFB with transcriptional activator GvpE. Arch Microbiol 194(4):269–279
84. Kixmuller D, Greie JC (2012) Construction and characterization of a gradually inducible expression vector for Halobacterium salinarum, based on the kdp promoter. Appl Environ Microbiol 78(7):2100–2105
85. Luttenberg B, Wolff EK, Engelhard M (1998) Heterologous coexpression of the blue light receptor psRII and its transducer pHtrII from Natronobacterium pharaonis in the Halobacterium salinarium strain Pho81/w restores negative phototaxis. FEBS Lett 426(1):117–120
86. Chizhov I, Schmies G, Seidel R, Sydor JR, Luttenberg B, Engelhard M (1998) The photophobic receptor from Natronobacterium pharaonis: temperature and pH dependencies of the photocycle of sensory rhodopsin II. Biophys J 75(2):999–1009
87. Kamekura M, Dyall-Smith ML, Upasani V, Ventosa A, Kates M (1997) Diversity of alkaliphilic halobacteria: proposals for transfer of Natronobacterium vacuolatum, Natronobacterium magadii, and Natronobacterium pharaonis to Halorubrum, Natrialba, and Natronomonas gen. nov., respectively, as Halorubrum vacuolatum comb. nov., Natrialba magadii comb. nov., and Natronomonas pharaonis comb. nov., respectively. Int J Syst Bacteriol 47(3):853–857
88. Bogomolni RA, Stoeckenius W, Szundi I, Perozo E, Olson KD, Spudich JL (1994) Removal of transducer HtrI allows electrogenic proton translocation by sensory rhodopsin I. Proc Natl Acad Sci U S A 91(21):10188–10192
89. Shand RF, Betlach MC (1991) Expression of the bop gene cluster of Halobacterium halobium is induced by low oxygen tension and by light. J Bacteriol 173(15):4692–4699
90. Ferrando-May E, Brustmann B, Oesterhelt D (1993) A C-terminal truncation results in high-level expression of the functional photoreceptor sensory rhodopsin I in the archaeon Halobacterium salinarium. Mol Microbiol 9(5):943–953
91. Khanna S, Srivastava AK (2005) Recent advances in microbial polyhydroxyalkanoates. Process Biochem 40(2):607–619
92. Han J, Lu Q, Zhou L, Zhou J, Xiang H (2007) Molecular characterization of the phaECHm genes, required for biosynthesis of poly(3-hydroxybutyrate) in the extremely halophilic archaeon Haloarcula marismortui. Appl Environ Microbiol 73(19):6058–6065
93. Kirk RG, Ginzburg M (1972) Ultrastructure of two species of halobacterium. J Ultrastruct Res 41(1):80–94
94. Li Z, Loh XJ (2015) Water soluble polyhydroxyalkanoates: future materials for therapeutic applications. Chem Soc Rev 44(10):2865–2879
95. Rehm BH, Steinbuchel A (1999) Biochemical and genetic analysis of PHA synthases and other proteins required for PHA synthesis. Int J Biol Macromol 25(1–3):3–19
96. Liebergesell M, Steinbuchel A (1992) Cloning and nucleotide sequences of genes relevant for biosynthesis of poly(3-hydroxybutyric acid) in Chromatium vinosum strain D. Eur J Biochem/FEBS 209(1):135–150
97. Cline SW, Lam WL, Charlebois RL, Schalkwyk LC, Doolittle WF (1989) Transformation methods for halophilic archaebacteria. Can J Microbiol 35(1):148–152

Part III

Lower Eukaryotic Expression Hosts

Production of Protein Complexes in Non-methylotrophic and Methylotrophic Yeasts

9

Nonmethylotrophic and Methylotrophic Yeasts

Francisco J. Fernández, Miguel López-Estepa, Javier Querol-García, and M. Cristina Vega

Abstract

Protein complexes can be produced in multimilligram quantities using nonmethylotrophic and methylotrophic yeasts such as *Saccharomyces cerevisiae* and *Komagataella* (*Pichia*) *pastoris*. Yeasts have distinct advantages as hosts for recombinant protein production owing to their cost efficiency, ease of cultivation and genetic manipulation, fast growth rates, capacity to introduce post-translational modifications, and high protein productivity (yield) of correctly folded protein products. Despite those advantages, yeasts have surprisingly lagged behind other eukaryotic hosts in their use for the production of multisubunit complexes. As our knowledge of the metabolic and genomic bottlenecks that yeast microorganisms face when overexpressing foreign proteins expands, new possibilities emerge for successfully engineering yeasts as superb expression hosts. In this chapter, we describe the current state of the art and discuss future possibilities for the development of yeast-based systems for the production of protein complexes.

Keywords

(Non-)methylotrophic yeast • *Saccharomyces cerevisiae* • *Kluyveromyces* • *Yarrowia* • *Komagataella* (*Pichia*) *pastoris* • *Ogataea* (*Hansenula*) *polymorpha*

F.J. Fernández (✉) • M. López-Estepa
J. Querol-García • M.C. Vega (✉)
Center for Biological Research, Spanish National Research Council (CIB-CSIC),
Ramiro de Maeztu 9, 28040 Madrid, Spain
e-mail: cvega@cib.csic.es

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_9

9.1 To Be or Not to Be (a Methylotroph)

A convenient phenotypical distinction between yeasts commonly used in protein expression is whether they can utilize methanol as a carbon and energy source or not. Those yeasts that can utilize methanol are generally termed "methylotrophic" yeasts, as opposed to yeasts that cannot thrive on methanol or "non-methylotrophic" yeasts [1]. Methylotrophic yeasts encode a complex network of specific enzymes tailored for the stepwise oxidation of methanol to formaldehyde and, ultimately, carbon dioxide, and typically rely predominantly or exclusively on oxidative respiration (*i.e.*, do not usually carry out anaerobic fermentation and therefore do not produce ethanol as a side product). Besides loosely coinciding with the phylogenetic placement of the two groups, this distinction is an operational convenience since protocols for cell growth and induction of recombinant expression within each group share significant similarities. In practice, the grouping is intended to evenly split yeasts in the two groups most often used for overexpression, epitomized by *S. cerevisiae* (non-methylotroph) *versus K. pastoris* (methylotroph).

9.2 Non-methylotrophic Yeasts

The group of non-methylotrophic yeast includes most prominently baker's yeast (*S. cerevisiae*), the budding yeast *Kluyveromyces lactis*, and the dimorphic yeast *Yarrowia lipolytica*. The wealth of knowledge on the microbiology [2], genetics [3], molecular and cellular biology [4], stress response [5, 6], and metabolism [7, 8] of *S. cerevisiae* is among the strongest factors in favor of its use for heterologous protein expression, especially on those cases where extensive optimization of the host cells is required for optimum yield and quality. This huge knowledge base allows targeted and global genetic engineering approaches as well as optimization of growth protocols and induction regimes in order to improve the yield of difficult eukaryotic proteins [5, 6, 9, 10]. However, one of its disadvantages as an expression host stems from its facultative anaerobic metabolism, which enables *S. cerevisiae* to undergo a metabolic or "diauxic" switch towards anaerobic consumption of glucose and the concomitant production of toxic ethanol, even in the presence of oxygen [11, 12]. *K. lactis* and the related species *K. marxianus* have many of the advantages of baker's yeast to which they add a marked preference for aerobic metabolism, which makes them attractive for recombinant expression. Finally, *Y. lipolytica* is a model organism for lipid catabolism and as such offers a number of advantages for the production of lipid-associated proteins and lipid metabolic complexes. Other non-methylotrophic yeasts (the fission yeast *Schizosaccharomyces pombe* and the dimorphic yeast *Arxula adeninivorans* are prime examples) have also been explored as recombinant hosts and are reviewed elsewhere [13].

9.2.1 *S. cerevisiae*

S. cerevisiae has long been a eukaryotic model organism of choice for a multitude of genetic, biochemical and metabolic processes, and has been used for the production of recombinant proteins [14]. A major advantage of *S. cerevisiae* as an expression host is that its genetics, metabolism and biochemistry are best known than those of any other yeast, and the number of different protein classes that have been studied in baker's yeast represents a huge knowledge base for future experiments. The applications for which *S. cerevisiae* has mainly exceled is the isolation of native complexes by (tandem) affinity purification [15] and the implementation of partial or complete metabolic pathways [16]. A schematic workflow of a typical overexpression experiment in *S. cerevisiae* is shown in Fig. 9.1.

A few coexpression plasmids have been published in the literature, all of them based on the GAL1/GAL10 bidirectional promoter: an autoselection based method [17, 18] and a two-plasmid co-transformation system allowing the simultaneous expression of up to four different genes.

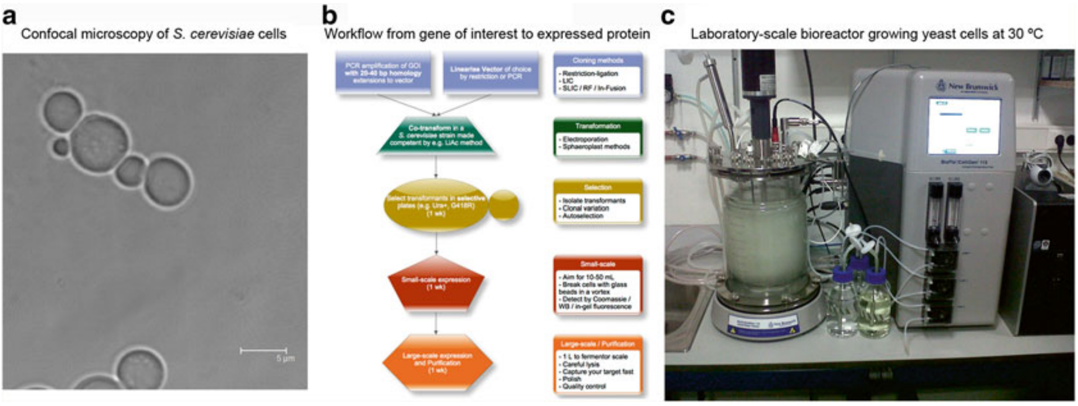

a
Confocal microscopy of S. cerevisiae cells
5 µm
b
Workflow from gene of interest to expressed protein
PCR amplification of GOI with 20-40 bp homology extensions to vector
Linearise Vector of choice by restriction or PCR
Cloning methods
- Restriction-ligation
- LIC
- SLIC / RF / In-Fusion
Co-transform in a S. cerevisiae strain made competent by e.g. LiAc method
Transformation
- Electroporation
- Sphaeroplast methods
Select transformants in selective plates (e.g. Ura+, G418R) (1 wk)
Selection
- Isolate transformants
- Clonal variation
- Autoselection
Small-scale expression (1 wk)
Small-scale
- Aim for 10-50 mL
- Break cells with glass beads in a vortex
- Detect by Coomassie / WB / in-gel fluorescence
Large-scale expression and Purification (1 wk)
Large-scale / Purification
- 1 L to fermentor scale
- Careful lysis
- Capture your target fast
- Polish
- Quality control
c
Laboratory-scale bioreactor growing yeast cells at 30 °C
New Brunswick

Fig. 9.1 Baker's yeast (*Saccharomyces cerevisiae*) as a protein expression platform. (**a**) Confocal microscopy image of growing *S. cerevisiae* cells, depicting the characteristic budding morphology of actively dividing yeast cells. (**b**) Schematic workflow from gene of interest (GOI) to expressed protein, assuming that the expression construct is assembled by *in vivo* homologous recombination cloning. Approximate times from preparation of DNA fragments to the evaluation of expression constructs generated according to the workflow at small scale and then at a larger scale range from 2 to 4 weeks

An example of a successful application of these systems for the production of multienzymatic systems consists of a bacterial seven-component isoprenoid biosynthetic pathway [19]. In either case, assembly of multigene constructs required multiple *ad hoc* cloning steps, such as the amplification of two independent GAL10/GAL1 expression cassettes and their subsequent restriction-ligation assembly into a single plasmid. Construction of yeast strains for more than four distinct genes requires, in these approaches, using at least two auxotrophic markers (*e.g., HIS3* and *TRP1*). For some applications, the strong repression exerted by glucose on expression from the GAL10/GAL1 promoters may be disadvantageous. In those cases, a strategy that has proven successful is to replace the GAL10/GAL1 tandem by a constitutively active TEF1/PGK1 promoter combination [20]. Most plasmids for co-production of proteins in *S. cerevisiae* have been devised for the assembly of multienzymatic pathways. Protocols often rely on the co-transformation of several such plasmids encoding different auxotrophic markers each of which drives the expression of one gene of interest using specific promoters. Those protocols are therefore time-consuming and the countless combinations of markers, genes and promoters can render the design of a typical coexpression experiment rather complicated. The combinatorial nature of this problem is exemplified by some recent proposed vector systems that ship with nearly 30 different plasmids [16, 21].

The highly efficient homologous recombination machinery of *S. cerevisiae* has been exploited for the in-vivo cloning of genes by supplying competent yeast cells with a linearized vector and one or several PCR fragments with terminal homologous sequences (>20 bp). The mixture of DNA fragments is repaired by *S. cerevisiae* cells and selective medium is used to retrieve only the correctly assembled plasmid. In this way, complete pathways can be constructed in essentially one step, and even small-sized whole genomes like that of *Mycoplasma genitalium* (592 kb) have been assembled in this way [22]. In a landmark study that further illustrates the power of homologous recombination, Annaluru et al. [23] succeeded in creating a completely redesigned *S. cerevisiae* chromosome III (616 Mb) in yeast cells. For enzymatic pathways, the same result can be achieved by using sequence- and ligation-independent cloning (SLIC) [24] rather than homologous recombination, as in the DNA assembler procedure [25]; using this method, functional D-xylose utilization and zeaxanthin biosynthesis pathways (8 genes, 19 kb in total) could be assembled.

An alternative strategy consists in engineering entry/recombination sites into the yeast's genome where multiple expression cassettes containing the genes of interest can be inserted sequentially; an elegant implementation of this strategy is the reiterative recombination method [26]. These strategies typically have the advantage that only one gene expression cassette is introduced at a time, hence a small number of auxotrophic or antibiotic markers can be efficiently recycled. Furthermore, the integration of the genes into the genome affords considerably greater stability to the recombinant strains in comparison with strains carrying a plasmid.

9.2.1.1 Examples

Preparation of the Mediator middle module [27].

The yeast Mediator middle module comprises seven different subunits. The med19Δ yeast strain (BY4741; *MATa*, *his3Δ1*, *leu2Δ0*, *met15Δ0*, *ura3Δ0*, *YBL093C*::*clonNAT*) was obtained from Gene Center Munich and C-terminal TAP-tags were introduced on Med7, Med15 or Med18, respectively, using a kanMX4-marker by means of vector pYM13 [28]. Briefly, kanMX4-marker and TAP-tags were amplified from pYM13 by PCR using primers containing within their 5′ region sequences of homology to the genomic target location; after that, cells were transformed with the resulting fragment and it was integrated into the yeast genome by homologous recombination. Yeast cultures were cultivated and the protein complexes purified using tandem affinity purification (TAP) [15]. The TAP tag consists of two IgG binding domains of *Staphylococcus aureus* protein A (ProtA) and a calmodulin binding peptide (CBP) separated by a TEV protease cleavage site. ProtA bound tightly to IgG Sepharose beads, requiring the use of the TEV protease to elute material under native conditions. The eluate of this first affinity purification step was then incubated with calmodulin-coated beads in the presence of calcium. After washing, which removed contaminants and the TEV protease remaining after the first affinity selection, the bound material was released under mild conditions with EGTA (Fig. 9.2). Mass spectrometry enabled to determine unambiguously that the endogenous yeast Mediator middle module comprised the subunits Med1, Med4, Med7, Med9 (Cse2), Med10 (Nut2), Med21 (Srb7) and Med31 (Soh1) (Fig. 9.2). In addition, it permitted to establish that only a single copy of each subunit is present in the complexes and the equimolar subunit stoichiometry of the middle module.

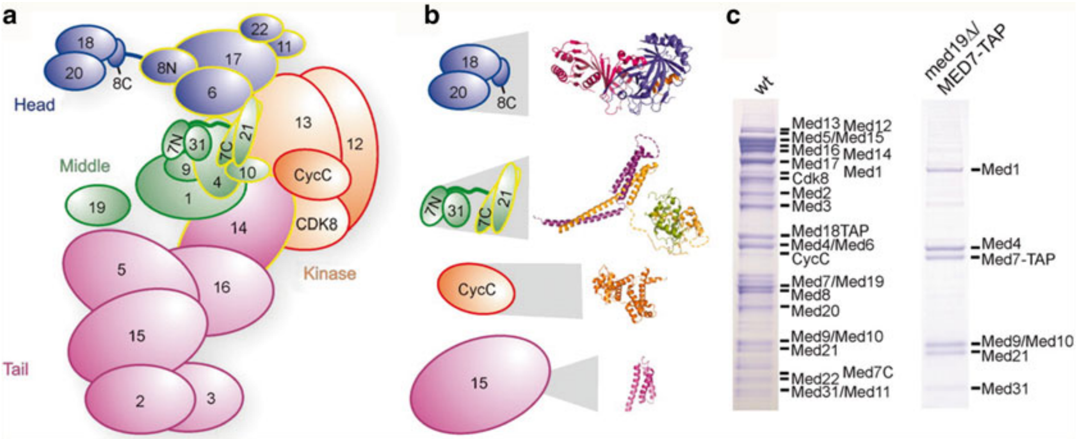


Fig. 9.2 Mediator subcomplexes produced in *S. cerevisiae*. A highlight example of using *S. cerevisiae* for the production of large multi-subunit complexes, expression of the Mediator complex. Schematic representation of the full Mediator complex (**a**) and of several subcomplexes (**b**). (**c**) Purified Mediator complexes analyzed by SDS-PAGE and stained with Coomassie Brilliant Blue. Subunits identified by mass spectrometry are indicated beside every band (Reproduced with kind permission from Koschubs et al. [27])

9.2.2 *K. lactis*

K. lactis [29, 30] and related yeast species like the thermotolerant *K. marxianus* [31] belong to the *S. cerevisiae* phylogenetic complex, and display numerous advantages for protein expression and biotechnology applications over baker's yeast. Most notable is the ability of *K. lactis* and *K. marxianus* to metabolize lactose as the sole carbon and energy source, which explains the wide use of these yeasts in the diary industry for the manufacture of milk, yoghurt and cheese products, including the native intracellular expression of lactase (β-galactosidase) for the production of lactose-free milk products [32]. Their rapid growth rate, aerobic metabolism and high secretory capacity are key properties for both yeasts as hosts for heterologous expression; the ability of *K. marxianus* to thrive up to 52 °C is an additional beneficial trait for industrial applications. The availability of strong constitutive/inducible promoters to drive high-level protein expression has fostered the commercialization of a *K. lactis* expression system (NEB) [29], while *K. marxianus* has being developed independently as a viable alternative.

The choice of yeast strain and genetic background is of great importance in the overall yield of an overexpression experiment. Two *K. lactis* strains are commonly used, strain CBS 2359 in academic settings and the food industry isolate GG799. *K. lactis* GG799 is a wild-type haploid strain which exhibits nearly no repression by glucose of the lactase (*LAC4*) promoter, therefore *LAC4*-driven expression is high using inexpensive medium compositions and glucose as a carbon source [29]. Various proteins have been successfully secreted by overproducing *K. lactis* cells, including human serum albumin (HSA), and these strains can be further improved by mutagenesis and by increasing the plasmid copy number, leading to "super-secreting" phenotypes.

Auxotrophic (*ura3*, *leu2*, *trp1*), dominant (*e.g.*, geneticin/G418) and nitrogen source selectable markers are available for isolating and maintaining *K. lactis* strains expressing the gene of interest. The nitrogen source selectable marker, based on the *A. nidulans* acetamidase gene (*amdS*), allows transformed *K. lactis* cells overexpressing acetamidase to obtain nitrogen from the breakdown of acetamide into ammonia. This approach works with wild-type strains as well as the diploid and aneuploid strains present in industrial strains subjected to extensive phenotypical selection. The success of this approach has been exemplified by the production of bovine enterokinase, ovalbumin, cellulose and mouse transthyretine [33], using the integrative expression vector pKLAC1. Acetamide selection can also be used to introduce iterative genetic modifications in prototrophic *K. lactis* strains since the marker can be recycled by counterselection with fluoroacetamide.

As in other yeasts, in *K. lactis* there is a choice between episomal and integrative vectors where there is a trade-off between the higher copy number attainable with episomal plasmids and the higher stability inherent to integrative plasmids. The copy number advantage of episomal plasmids can be compensated for integrative plasmids by methods that increase the naturally small chance (2–5 %) of multiple copy integrations. For example, using pKLAC1 combined with acetamide selection (instead of G418) strains can be isolated that have integrated 2–6 copies in >90 % of the transformed cells. Other strategies involve expressing the selectable marker from a weak or attenuated promoter, or directing plasmid integration to chromosomal regions where recombination events are frequent.

Expression of foreign proteins in *K. lactis* is often driven by *S. cerevisiae* or *K. lactis* constitutive and regulatable promoters, including the glycolytic *S. cerevisiae PGK1* promoter, the phosphate inducible acid phosphatase (*S. cerevisiae PHO5*) promoter, and the galactose/lactose inducible *K. lactis LAC4* promoter. None of these promoters is fully repressed even in the absence of the inducer, potentially causing problems when preparing constructs in *E. coli*. A recently described *LAC4-BI* promoter variant circumvents this problem by deleting sequences akin to the bacterial Pribnow box transcriptional element [33], leading to a fully inhibited promoter in *E. coli* without negatively affecting expression in

K. lactis. Many other promoters have been tested in *K. marxianus*, including the strong constitutive *TDH3* promoter and the exogenous, tetracycline-repressible "tet-off" promoter.

The reiterative use of acetamidase selection/counterselection provides a straightforward method for the construction of *K. lactis* coexpression strains, either multi-copy single-protein constructs or multi-subunit protein complex constructs. Each selection/counterselection step inserts one expression cassette and eliminates the *amdS* gene, thus allowing the introduction of one more gene and its positive selection in agar plates supplemented with acetamide. Using this approach, up to four different plasmids have been simultaneously introduced in *K. lactis* cells and proven to yield stable expression of all four recombinant genes (Fig. 9.3) [34]. Co-production strains can also be generated using the same reiterative scheme with the aid of alternative recycling markers, such as the "*URA3* blaster" or the Cre-loxP systems [35].

A more traditional approach was employed by Hong et al. [36] to simultaneously co-express three cellulose degrading enzymes, endo-β-1,4-glucanase (*eng1* gene), cellobiohydrolase (*cbh1* gene) and cellulase (*bgl1* gene), inserted in *K. marxianus* chromosome. In this approach, a *K. marxianus* auxotrophic strain for the *URA3* locus was created by gene disruption using a selectable KanMX cassette, and then, this strain was used to introduce the *LEU2* and *TRP1* auxotrophies by the *URA3* blaster method, thus generating a triple auxotrophic strain; a final round of counterselection with FOA liberated the recyclable *hisG-URA3-hisG* marker. Exploiting the targeting sites created in the auxotrophic strain, several cellulose-degrading strains were constructed including a strain expressing the three genes under the control of the GAP or ADH promoters into two distinct loci (*URA3* and *bla*) [*URA3*::*GAP-eng1-ADH-cbh1* *bla*::*GAP-bgl1*] and a strain expressing five gene copies in four loci (*rDNA*, *LEU2*, *TRP1* and *URA3*) [*rDNA*::*GAP*-α-*eng1* *LEU2*::*GAP*-α-*eng1* *bla*::*GAP-bgl1* *URA3*::*GAP*-α-*eng1-ADH-cbh1*]. These strains did not only overexpress the three cellulolytic enzymes but, owing to their thermostability and the ability of *K. marxianus* to grow at relatively elevated temperatures, made possible the efficient bioconversion of cellobiose into ethanol at 45 °C.

In a different strategy to engineer a coexpressing *K. lactis* strain, the two genes encoding the large and small subunits of the atypical heterodimeric POXA3 laccase from the white-rot fungus *Pleurotus ostreatus* were cloned in a modified pKD1 expression vector bearing a bidi-

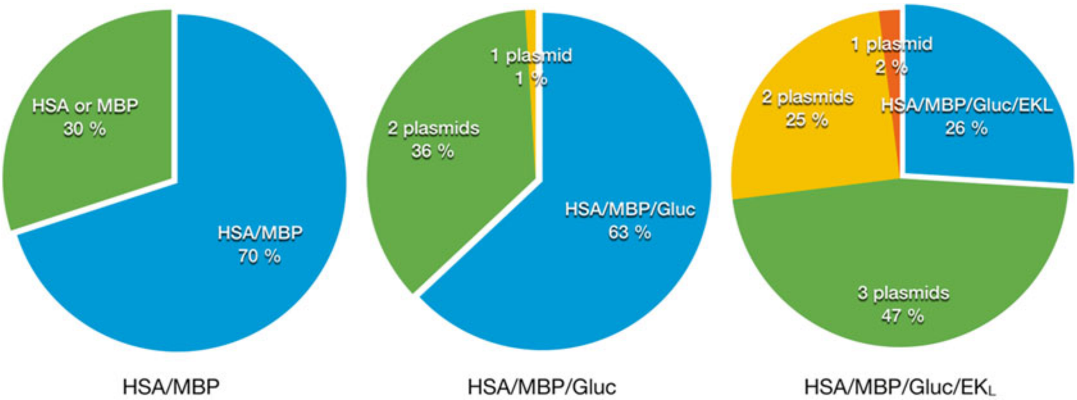


Fig. 9.3 Multiple plasmid co-transformation in *K. lactis*. Acetamide selection of *K. lactis* cells transformed with expression plasmids is very efficient thereby allowing the simultaneous transformation with multiple expression constructs [34]. Pie chart representation showing the frequencies of co-selected expression plasmids after co-transforming with two, three, or four different plasmids encoding human serum albumin (HSA), *Escherichia coli* maltose binding protein (MBP), *Gaussia princeps* luciferase (Gluc) or bovine enterokinase light chain (EKL), which allowed the direct preparation of multi-protein expressing strains growing in acetamide (Modified from Read et al. [34])

rectional promoter [37], including their native secretion leader sequences. The pKD1 plasmid is the only natural plasmid found in *Kluyveromyces*, it was first isolated from *K. drosophilarum* and is maintained at 60–80 copies per cell for many generations [38]. Each POXA3 gene was cloned under the control of the endogenous bidirectional *KlADH4* promoter by using restriction/ligation methods. The expression construct was transformed into the *K. lactis* CMK5 strain and plasmid-containing cells were isolated on selection medium supplemented with G418. Addition of 0.5 % ethanol per day was used to trigger recombinant expression of the two-subunit POXA3 laccase, with 1 mM $CuSO_4$ supplementation for direct incorporation into the expressed POXA3. Analysis of secreted protein indicated that functional POXA3 started to accumulate in the culture medium from day 5 onwards, reaching the maximum laccase activity around 14–17th day. In accord with the proposed role of POXA3 small subunit in stabilizing the catalytic large subunit, expression of only the large subunit in an otherwise identical setting led to the intracellular accumulation of inactive enzyme; in contrast, simultaneous coexpression of both the large and small POXA3 subunits yielded far more secreted and functional heterodimeric POXA3 complex [37].

As explained above, using acetamide-based selection in nitrogen-free medium (as implemented in the pKLAC1 vector) automatically leads to multiple insertions up to a total of 2–6 integrated copies of the expression cassette. This can be exploited to increase the gene dosage of the target protein or, more interestingly, for the coexpression of several different proteins [34]. Protypical examples are the coexpression *K. lactis* strains generated by co-transformation with pKLAC1 plasmids encoding the light and heavy chain Fab fragments of anti-MBP and anti-ferritin antibodies preceded by the α-factor secretion leader [34]. Although the co-transformation method has been successfully applied, the dose of each co-transformed open reading frame is uncertain and bound to vary from colony to colony, requiring extensive screening to identify suitable expression strains. The method is particularly useful as a screening tool when downstream analysis of successful expression tests can be determined easily by a color/fluorescence or a functional assay (*e.g.*, an enzymatic assay) that can be applied in microtiter format. When the stoichiometry of the expressed protein chains is crucial for the successful assembly and/or function of a protein complex, strategies where a multigene construct is cloned may be preferable to co-transformation.

9.2.3 *Y. lipolytica*

A recent review has been published on the properties of *Yarrowia lipolytica* as a host for expression and secretion of heterologous proteins [39]. Here, we will focus on those properties of *Y. lipolytica* that are particularly relevant for the production of protein complexes.

Y. lipolytica [40] is a heterothallic, hemiascomycete yeast with a sexual reproduction mechanism characterized by a haplo-diplontic cycle, with one mating-type locus and two alleles, MATA and MATB, thereby allowing standard genetic manipulations to be performed. *Y. lipolytica* is phylogenetically very distantly related to the *Saccharomyces* clade, thereby it presumably offers distinct advantages and opportunities from the more closely related *Saccharomyces*/*Schizosaccharomyces*/*Kluyveromyces* yeasts as a host for heterologous expression. With a 49 % G+C ratio, *Y. lipolytica* codon bias is closer to that of *Aspergillus*, and it shares other genomic properties with filamentous fungi [41]. This yeast has been employed as a model system for yeast genetics and, in particular, for the study of lipid biosynthesis and catabolism, mainly because it can grow on alkane sources, oleic acid, polyalcohols, organic acids or paraffin [42]. This particular growth capacity makes *Y. lipolytica* an attractive yeast for the expression of enzymes involved in lipid metabolism, many of which are overexpressed in yeast cells growing on fatty acids. For the same reason, promoters that drive the expression of fatty acid metabolic enzymes tend to be strong and therefore suitable for the heterologous production of proteins. Another promoter that is suitable for the overexpression of proteins is the *XPR2* promoter [40, 43], which in wild-type *Y. lipolytica* strains drives the high-level secretion of an extracellular alkaline protease (up to 1–2 g/L medium). Bovine prochymosin [44], the human anaphylatoxin C5a [45] and tissue plasminogen are three natively extracellular proteins that have been successfully produced in *Y. lipolytica* under the control of the *XPR2* promoter.

Very few examples exist of coexpression constructs cloned into *Y. lipolytica*. A recent one consists in the simultaneous expression of two fatty acid desaturase (PUFA) genes from *M. alpina*, Δ12-desaturase and Δ6-desaturase genes. The two genes were cloned into the same plasmid under the control of the strong and quasi-constitutive hybrid hp4d promoter [46], and inserted into *Y. lipolytica* genome by transformation with the linearized plasmid. The coexpression strain was isolated and proved to be efficient in producing both enzymes simultaneously in an efficient manner, what was further established by examining the fatty acid profile of the overexpression strain, where measurable amounts of the end product ϒ-linolenic acid (GLA) accumulated [47]. Using this coexpression strain as starting point, Chuang et al. [47] already discuss that other fatty acid biosynthetic enzymes could be added to their coexpression construct to create further diversity of polyunsaturated fatty acids, including the genes for elongase and Δ5-desaturase.

One of the topics where improvement of *Y. lipolytica* directly depends on the coproduction of proteins: in contrast to *S. cerevisiae* and *K. pastoris*, where strains have been constructed that express chaperones and foldases to support the correct folding of heterologous proteins, in *Y. lipolytica* the instability of the available plasmids has hampered progress in this area [39]. A pioneering example is the coexpression of the endoproteinase Xpr6p and an scFv fragment cloned downstream of the *XPR2* pre-pro sequence, which demonstrated the increase in mature secreted scFv by the coexpression strain in comparison with a strain expressing the scFv gene only [46].

9.3 Methylotrophic Yeasts

Methylotrophic yeasts are phenotypically characterized by being able to feed in methanol as the sole source of energy and carbon atoms. Two methylotrophic yeasts have been widely used, the mesophilic *K. pastoris* and the thermotolerant *Hansenula* (*Ogataea*) *polymorpha*. Both microorganisms have found wide acceptance in the academic and the industrial communities. *K. pastoris*, in particular, has rapidly established as one of the preferred yeasts for recombinant overexpression of proteins and has been extensively reviewed elsewhere in the context of the production of heterologous proteins in general [48] as well as for applications for the food and feed industries [49]. Here, we focus on the use of *K. pastoris* and *O. polymorpha* as hosts for the production of multisubunit protein complexes.

9.3.1 *K. pastoris*

One of the most versatile yeast microorganisms for recombinant protein production is *K. pastoris*. Most expression vectors commonly used depend on the very strong, highly regulated alcohol oxidase 1 (AOX1) promoter, which is switched on in presence of methanol and absence of other alternative carbon sources (glycerol and glucose). The enormous protein productivity per cell of *K. pastoris* motivated the proposal by Phillips Petroleum Company and the Salk Institute that *K. pastoris* could be used as a cheap source of protein for human and animal consumption (single cell protein, SCP) [50]. Although the oil crisis of the 1970s truncated that plan, *K. pastoris* has developed as a viable host for recombinant protein production and as a model yeast system for peroxisome research. Recently, *K. pastoris* has been used as a whole-cell vaccine (as heat-killed yeast cells) [51].

K. pastoris shares with other yeasts some desirable properties for protein overexpression: lower costs for media components than other eukaryotic hosts, shorter generation times, and introduction of post-translational modifications (disulfide bond, glycosylation, proline cis/trans isomerization, disulfide isomerization, lipidation, sulfation and phosphorylation) [52]. *K. pastoris* is especially well suited for the secretion of recombinant proteins, due to the high efficiency of its secretory pathway, which also permits the recovery of the secreted proteins from the extracellular medium with ease and in high purity. In comparison with *S. cerevisiae*, *K. pastoris* does not generate appreciable amounts of ethanol as metabolic by-product, which reduces toxicity and allows *K. pastoris* to grow to higher cell densities than alternative yeast hosts. Since *K. pastoris* can efficiently grow in mineral defined (minimal) medium, it has been used for the production of isotopically (^{15}N and ^{13}C) labeled proteins [53]. The availability of *K. pastoris* strains knocked out for vacuolar proteases, as SMD1168H, which lacks the vacuolar peptidase A gene, increases the stability of protease-sensitive protein products during expression and lysis [54].

The secret behind the great success of *K. pastoris* in recombinant protein production lies in the strength and highly regulated nature of the AOX1 promoter [50], which is the best understood promoter in *K. pastoris*. The *AOX1* gene product is involved in the first step of methanol catabolism. The strength of the AOX1 promoter is so great that the protein product can reach up to 30 % of the cell total protein [52]. In *K. pastoris* there is a second alcohol oxidase promoter, AOX2, which is weaker than AOX1 and has a failsafe function. *K. pastoris* strains that express the AOX1 gene product display a phenotype designated as Mut$^+$ (methanol utilization plus). Strains where the *AOX1* gene is knocked out or disrupted by nonsense mutations rely exclusively on the weaker AOX2 promoter for methanol utilization and grow more slowly; the resultant phenotype is Mut^S (methanol utilization slow). When both the *AOX1* and *AOX2* genes are disrupted, *K. pastoris* cannot grow only on methanol and the corresponding phenotype is Mut$^-$ [55]. There are certain cases, *e.g.*, horse radish peroxidase (HRP), where the slower translation rates assumed to occur in Mut$^-$ strains result in higher expression yields, presumably because of enhanced folding [55].

Besides the AOX1 promoter, there are other promoters available for use in *K. pastoris*. The FLD1 (formaldehyde dehydrogenase 1) promoter drives the expression of another gene in the methanol utilization pathway [56]. The FLD1 promoter is induced by both methanol and methylamine [57]. An example of the successful use of the FLD1 promoter is the expression of *R. oryzae* lipase [56]. An alternative to the AOX1 and FLD1 promoters is the constitutive glycolytic promoters, like the glyceraldehyde-3-phosphate dehydrogenase (GAP) promoter. An obvious argument in favor of the GAP promoter is that methanol is not required, thereby reducing toxicity and fire hazard considerations. In some cases, expression levels were higher using the GAP promoter than with the AOX1 promoter, *e.g.*, β-lactamase [58]. One further example of a promoter successfully used in *K. pastoris* is that of ICL1 (isocitrate lyase), which is inducible by ethanol in the absence of glucose [59]. An example of the application of the ICL1 promoter is the expression of the dextranase from *Penicillium minioluteum* [59].

The pattern of glycosylation of recombinant proteins secreted by *K. pastoris* of recombinant proteins is an important consideration for proteins of mammalian origin. In comparison with other yeasts like *S. cerevisiae*, *K. pastoris* glycans are shorter in chain length and have fewer branches, which makes *K. pastoris* more suitable as a host for the production of glycosylated products [60]. For some applications, though, correctly folded, soluble glycosylated protein is not sufficient—for human applications, the glycosylation pattern must approximate an authentic mammalian-type glycosylation [61]. The possibility of engineering the glycosylation pathway in *K. pastoris* has been exploited to generate yeast strains that mimic mammalian glycosylation patterns, which can aid the expression of heterologous proteins that are known to be glycosylated by their native hosts [62].

9.3.1.1 General Procedures

Linearized expression plasmids are typically transformed in *K. pastoris* by electroporation, and cells that have successfully integrated the expression cassette by homologous recombination into the desired locus (*e.g.*, AOX1, GAP, or randomly) are selected for by plating the transformation mixture onto Luria-Bertani plates supplemented with a suitable antibiotic. Zeocin is the most widely used antibiotic, but G418, blasticidin, and others, have also been used. Alternatively, auxotrophic markers can be used, the most popular ones being *HIS4* and *ADE2*. Genome integration is affected by a certain degree of heterogeneity; hence, selection of successful integrands is an important step. Of the many parameters that may potentially affect integration, the length of the linearized plasmid (hence, the length of the gene of interest) is the most critical variable. Over all possible transformation methods, electroporation after a pretreatment with lithium acetate and dithiothreitol (DTT) seems to be the most successful [63], achieving transformation efficiencies in the 10^6 cfu/μg, similar to efficiencies reported for *S. cerevisiae* or *E. coli*.

Multiple integration events are possible in *K. pastoris* leading to more copies of the gene of interest becoming integrated in the chromosome, with the expected result of higher expression yields with more integrated copies [64]. There are also cases where a greater gene copy number translated into *less* expressed product [65], therefore screening for the optimum, rather than the maximum, number of integrated gene copies is highly advisable. Convenient techniques to experimentally determine the number of integrated gene copies include Southern blotting [66] and RT-PCR [67].

K. pastoris secretory pathway is highly effective at channeling expressed proteins into the extracellular medium. Foreign proteins can be targeted for secretion by fusing the gene of interest with a suitable peptide signal. The most widely used is the α-mating factor secretion signal from *S. cerevisiae* [54], which is cleaved internally by the Kex2 [68] and, optionally, Ste13 proteases. Other secretory signal peptides have been tried, such as the α-melting factor with HRP and lipases, attaining up to twofold improvements in the yield of secreted product [68], and the acid phosphatase signal (PHOI) from *K. pastoris*, which was employed for glucoamylase [69] and NOP-1 [70].

The ability of *K. pastoris* to grow to very high cell densities makes it an interesting microorganism for the industrial production of proteins at large scale. In particular, *K. pastoris* grows very well in bioreactors on minimal defined medium and the high stability of the integrated expression cassettes reduces production costs by eliminating the need to use antibiotics. Numerous growth parameters can be monitored and tuned to increase yield, *e.g.*, aeration, pH, temperature, cell density, and induction regime. *K. pastoris* is relatively insensitive to the pH of the cultivation medium and it thrives from pH 3 to 7, although it must be borne in mind that certain vacuolar proteases that are liberated to the medium by spontaneous cell lysis are activated at acidic pH [65]. Various strategies are used to reduce the harmful effect of proteases on the expressed product, including addition of casamino acids or peptone to the medium [71].

Expression in *K. pastoris* can be performed in conventional shake flasks or in bioreactors. If induction is triggered by methanol, as for the AOX1 promoter, sufficient supply of oxygen gas is necessary to sustain expression since *K. pastoris* is a strict aerobe and consumes large amounts of O_2 during expression [72].

One area in which *K. pastoris* has had a great impact is in the production of eukaryotic membrane proteins, *e.g.*, for crystallography or drug discovery. For example, transporters such as CTR1 (human copper transporter) [73], ion channels such as the voltage-dependent K^+ channel [74], or GPCRs such as adenosine A_{2A} receptor [75]. Since *K. pastoris* membranes lack cholesterol, its supplementation during expression may be necessary for the correct folding and insertion of the overexpressed membrane proteins [76]. A more robust strategy was implemented for the expression of the Na^+, K^+-ATPase α3β1 isoform, where *K. pastoris* cells were engineered to express cholesterol in place of the natural ergosterol [77].

9.3.1.2 Examples

Whole *K. pastoris* yeast expression of measles virus nucleoproteins for production and delivery of multimerized *Plasmodium* antigens [51].

Vaccine development and deployment is often hampered by the unsuitability of effective and safe adjuvants. *K. pastoris* has been shown to be a useful expression platform for the production of vaccine antigens and as a novel delivery system, whereby whole recombinant yeast cells that have expressed the desired antigen are heat inactivated and injected to mice. A convincing proof of concept was presented for the circumsporozoite protein (CS) from *Plasmodium*, the etiologic agent of malaria. *P. berghei* CS (PbCS) was multimerized by fusion with the measles virus (MV) nucleoprotein (N), which is known to assemble into large-size ribonucleoprotein rods (RNPs) in yeast. Expression of the fusion N-PbCS protein in *K. pastoris* yielded large amounts of RNPs that could be analyzed by electron microscopy and immunofluorescence (Fig. 9.4). RNPs were shown to localize in peripheral cytoplasmic inclusions. Remarkably, subcutaneous immunization of a malaria infection mice model revealed that even a modest amount of yeast-borne N-PbCS RNPs elicited a reduction in parasitemia after a high dose of parasites had been administered. This study highlights the usefulness of heat-inactivated *K. pastoris* expressing N-PbCS RNPs as a source of new and cheap vaccine-adjuvant formulations.

Co-expression of human insulin-like growth factor II (IGF-II) and insulin-like growth factor binding protein 6 (IGFBP-6) [78].

The most common strategy to create coexpression constructs for *K. pastoris* is to clone each individual gene into expression plasmids and then use restriction-ligation techniques to concatenate the expression cassettes. This is laborious and it can become inefficient with large genes, where finding restriction enzymes that do not cut into the genes becomes problematic. In this case, 2×IGFBP-6 and 2×IGF-II were introduced separately into the pAO815 vector fused 3′ to the α-factor signal. Then, the double digestion product from pAO815-2×IGFBP-6 with BamHI-BglII was ligated into dephosphorylated pAO815-2×IGF-II, thereby generating the pAO815-2×IGF-II-2×IGFBP-6 plasmid for secretion. The plasmid was electroporated into *K. pastoris* GS115 cells after linearization with

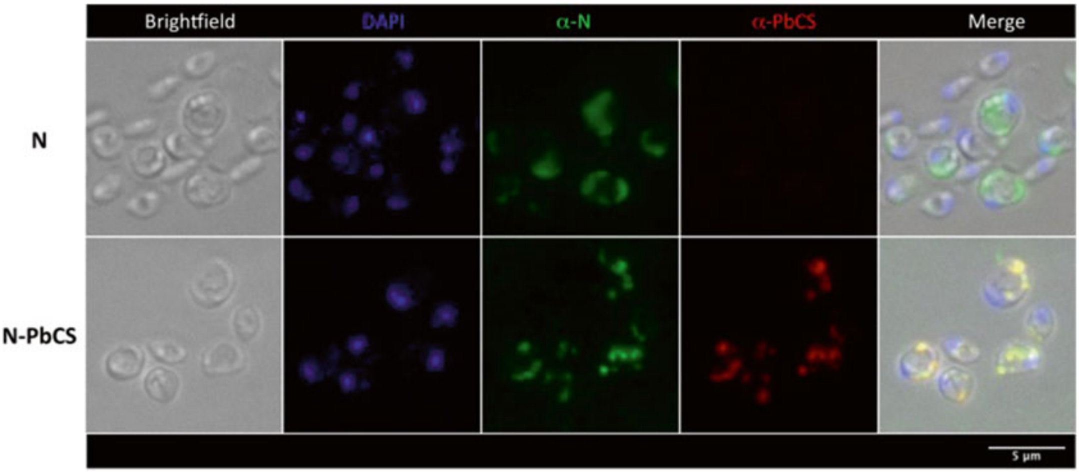


Fig. 9.4 Multimerized N-PbCS RNPs for vaccine development expressed in *K. pastoris. K. pastoris* was used to express the measles nucleoprotein (N) and a fusion of N and *Plasmodium berghei* circumsporozoite protein, PbCS (N-PbCS). Immunofluorescence (N staining in *green*, PbCS staining in *red*, nuclei in *blue*) was used to determine the localization of the multimerized N-PbCS RNPs. Conventional brightfield confocal microscopy and fluorescence microscopy were used to investigate the intracellular location of each subunit, thereby revealing a functional pattern of spatial distribution for the complex subunits. The combined image (*right*) further illustrates the spatial coincidence of the overexpressed proteins (reproduced with kind permission from Jacob et al. [51])

SalI, and recombinant clones were selected on MD medium without histidine. Coexpression and secretion of correctly folded IGFBP-6-IGF-II complex was accomplished by methanol induction (1 %) at 30 °C.

Co-expression of recombinant human Claudin-1 and CD81 [79, 80].

The Claudin-1-CD81 complex is representative of the complexes formed by members of the claudin and tetraspanin superfamilies. A structure of the Claudin-1-CD81 complex, still unavailable, would promote the development of therapeutic agents targeting the early entry step of the human cytomegalovirus (HCV) lifecycle. The CD81 gene was amplified using a mutagenic PCR strategy designed to exchange all palmitoylation sites (Cys to Ala) and insert a C-terminal hexahistidine tag, and cloned into pPICZB for intracellular expression. The construct was linearized and transformed in *K. pastoris* X33 and GS115 cells by electroporation, and positive transformants were selected by zeocin resistance. Clones expressing CD81 were screened at 30 °C inducing with 1 % (w/v) methanol over 48 h, and solubilization and purification of CD81 was conducted using suitable detergents: 2 % (w/v) n-octyl-β-D-glucopyranoside (β-OG), lauryldimethylamine oxide (LDAO) or n-dodecylphosphocholine–cholesterolhemisuccinate (DPC/CHS). In turn, Claudin-1 was solubilized using 3 % (w/v) β-OG, profoldin-8 or foscholine-10 (FC10).

9.3.1.3 *O. polymorpha*

The second most popular methylotrophic yeast for recombinant protein production is the thermotolerant *Ogatella* (*Hansenula*) *polymorpha*. It is phylogenetically close to *K. pastoris* and other methylotrophic yeasts, with which *O. polymorpha* shares the methanol utilization pathway, while it sets apart from all others by also having a pathway for nitrate assimilation. A second characteristic unique to *O. polymorpha* is its capacity to withstand temperatures between 30 and 49 °C, which makes it attractive for industrial applications where bioprocess temperature might be tuned up for the expression of thermostable enzymes. In analogy to *K. pastoris*, the strongest promoters in *O. polymorpha* have been derived from the methanol utilization pathway. Two such promoters, both inducible by methanol, have been derived from the methanol oxidase (*MOX*) and formate dehydrogenase (*FMD*) genes [81]. In contrast to the *AOX1* promoter from *K. pastoris*, the *MOX* promoter, which is tightly regulated in rich medium, becomes significantly derepressed under glycerol starvation and therefore does not necessarily require methanol for induction (although methanol may be added during the last phase of cultivation). More often, expression of the heterologous protein is induced by switching to a glycerol limiting feed, which is optimum if ramped up in correlation to the cell mass [82]. An alternative to the methanol-inducible promoters in *O. polymorpha* is the *TPS1* (trehalose-6-phosphate synthase) promoter, which is inducible through the heat-shock response by culturing expressing strains at 37–40 °C, and was shown to drive expression of two reporter genes (*lacZ* and *GFP*) to very high titers [83]. The value of *O. polymorpha* for the production of heterologous proteins is well established, especially in industrial settings.

Successfully produced proteins include interferon IFNα-2a [84], interleukin-6 (IL-6) [85], phytase [82] and the heat shock protein gp96 [86]. As early as 1996 *O. polymorpha* was shown to perform well as a host for coexpression of heterologous proteins, when a recombinant strain was engineered for the production of glyoxylic acid from glycolate with the concomitant decomposition of the product peroxide into water and oxygen [87]. Two genes were inserted into *O. polymorpha* chromosome to accomplish this: spinach glycolate oxidase (GO) and *S. cerevisiae* catalase T (CTT1) genes.

The highly clustered nature of the ribosomal RNA (rRNA) genes has been utilized for the single-step co-integration of several expression cassettes into *O. polymorpha* genome. In a pioneering study, up to four expression vectors encoding several reporter proteins (GFP, lacZ, insulin and phytase) were co-transformed using a single auxotrophic marker (*URA3*), and the stable integration of the cassettes and the expression of the target genes were confirmed [88]. This is an interesting approach for evaluating coexpression strains, especially during the initial screening phase, where well performing candidate strains can be selected for further analysis.

Coexpression of target proteins with chaperones and foldases has been applied to *O. polymorpha*. The endogenous gene for calnexin (*OpCNE1*) encodes a 557-amino-acid membrane protein that appears to monitor the folding state of nascent polypeptides as they are translocated into the endoplasmic reticulum. When *Op*CNE1 was overexpressed in strains that also secreted a second target protein (interferon-γ, AlgE1 and a consensus phytase) the overall yield of secreted protein product increased up to fivefold.

9.4 Conclusions

Unicellular yeast microorganisms offer many advantages for the recombinant production of protein and protein complexes over conventional bacterial hosts, namely: improved folding properties, post-translational modifications (*e.g.*, glycosylation), target localization (*e.g.*, secretion to the extracellular medium), growth to very high cell densities and high productivity per cell. The cost-efficiency and fast growth rates compares favorably with other eukaryotic expression systems like insect and mammalian cells. Although ultimately the usefulness of using a yeast-based expression system over the alternative systems might depend on a balanced compromise between these various factors, the substantial body of knowledge accumulated on yeast genomics and proteomics and the ease with which new strains can be engineered certainly present yeasts as attractive hosts for heterologous production of macromolecular complexes. Tapping into the variety of yeast species that are available will become easier in the future thanks to initiatives to build plasmids which can drive heterologous expression of proteins and protein complexes with one plasmid only without further modifications, since then the identity of the yeast host becomes a screening variable that is optimized for each target recombinant protein of choice (Fig. 9.5). This and new exciting opportunities will surely become available as the huge pool of knowledge on yeast microorganisms is exploited more systematically.

Acknowledgments This work was funded by the Spanish Ministry of Economy and Competitiveness grant CTQ2015-66206-C2-2-R and the European Commission (Framework Programme 7) (FP7) project ComplexINC No. 279039 to M.C.V.

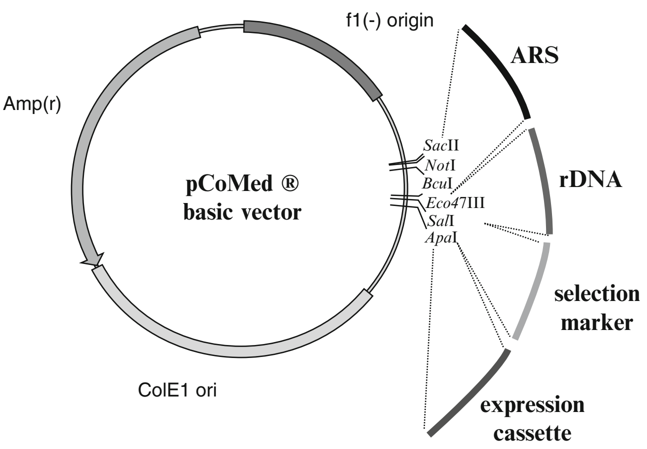


Fig. 9.5 Beyond the one host-one plasmid paradigm for yeast. Exploitation of the specific properties of yeast species is a rich source of novel hosts for heterologous protein expression. However, phylogenetically distinct yeasts share common genetic and metabolic traits that open the way to the development of plasmids that operate successfully across yeast species. The CoMed system, illustrated herein, is one such system that performs protein expression in nonmethylotrophic and methylotrophic yeasts including all yeasts discussed in this chapter (except *K. lactis*) (reproduced with kind permission from Gellisen et al. [89])

References

1. Fernandez FJ, Vega MC (2013) Technologies to keep an eye on: alternative hosts for protein production in structural biology. Curr Opin Struct Biol 23(3):365–373
2. Barnett JA, Barnett L (2011) Yeast research: a historical overview. ASM Press, Washington, DC
3. Duina AA, Miller ME, Keeney JB (2014) Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. Genetics 197(1):33–48
4. Resnick MA, Cox BS (2000) Yeast as an honorary mammal. Mutat Res 451(1–2):1–11
5. Bawa Z, Bland CE, Bonander N, Bora N, Cartwright SP, Clare M, Conner MT, Darby RA, Dilworth MV, Holmes WJ, Jamshad M, Routledge SJ, Gross SR, Bill RM (2011) Understanding the yeast host cell response to recombinant membrane protein production. Biochem Soc Trans 39(3):719–723
6. Bonander N, Bill RM (2012) Optimising yeast as a host for recombinant protein production (review). Methods Mol Biol 866:1–9
7. Barnett JA (2003) A history of research on yeasts 5: the fermentation pathway. Yeast 20(6):509–543
8. Barnett JA (2003) A history of research on yeasts 6: the main respiratory pathway. Yeast 20(12):1015–1044
9. Ashe MP, Bill RM (2011) Mapping the yeast host cell response to recombinant membrane protein production: relieving the biological bottlenecks. Biotechnol J 6(6):707–714
10. Bill RM (2014) Playing catch-up with Escherichia coli: using yeast to increase success rates in recombinant protein production experiments. Front Microbiol 5:85
11. Verduyn C, Zomerdijk TL, van Dijken J, Scheffers WA (1984) Continuous measurement of ethanol production by aerobic yeast suspensions with an enzyme electrode. Appl Microbiol Biotechnol 19(3):181–185
12. Darby RA, Cartwright SP, Dilworth MV, Bill RM (2012) Which yeast species shall I choose? Saccharomyces cerevisiae versus Pichia pastoris (review). Methods Mol Biol 866:11–23
13. Celik E, Calik P (2012) Production of recombinant proteins by yeast cells. Biotechnol Adv 30(5):1108–1118
14. Romanos MA, Scorer CA, Clare JJ (1992) Foreign gene expression in yeast: a review. Yeast 8(6):423–488
15. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods 24(3):218–229
16. Fang F, Salmon K, Shen MW, Aeling KA, Ito E, Irwin B, Tran UP, Hatfield GW, Da Silva NA, Sandmeyer S (2011) A vector set for systematic metabolic engineering in Saccharomyces cerevisiae. Yeast 28(2):123–136
17. Geymonat M, Spanos A, Sedgwick S (2009) Production of mitotic regulators using an autoselection system for protein expression in budding yeast. Methods Mol Biol 545:63–80
18. Geymonat M, Spanos A, Sedgwick SG (2007) A Saccharomyces cerevisiae autoselection system for optimised recombinant protein expression. Gene 399(2):120–128
19. Maury J, Asadollahi MA, Moller K, Schalk M, Clark A, Formenti LR, Nielsen J (2008) Reconstruction of a bacterial isoprenoid biosynthetic pathway in Saccharomyces cerevisiae. FEBS Lett 582(29):4032–4038
20. Partow S, Siewers V, Bjorn S, Nielsen J, Maury J (2010) Characterization of different promoters for designing a new expression vector in Saccharomyces cerevisiae. Yeast 27(11):955–964
21. Shen MW, Fang F, Sandmeyer S, Da Silva NA (2012) Development and characterization of a vector set with regulated promoters for systematic metabolic engineering in Saccharomyces cerevisiae. Yeast 29(12):495–503
22. Gibson DG, Benders GA, Axelrod KC, Zaveri J, Algire MA, Moodie M, Montague MG, Venter JC, Smith HO, Hutchison CA 3rd (2008) One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. Proc Natl Acad Sci U S A 105(51):20404–20409
23. Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, Richardson SM, Dymond JS, Kuang Z, Scheifele LZ, Cooper EM, Cai Y, Zeller K, Agmon N, Han JS, Hadjithomas M, Tullman J, Caravelli K, Cirelli K, Guo Z, London V, Yeluru A, Murugan S, Kandavelou K, Agier N, Fischer G, Yang K, Martin JA, Bilgel M, Bohutski P, Boulier KM, Capaldo BJ, Chang J, Charoen K, Choi WJ, Deng P, DiCarlo JE, Doong J, Dunn J, Feinberg JI, Fernandez C, Floria CE, Gladowski D, Hadidi P, Ishizuka I, Jabbari J, Lau CY, Lee PA, Li S, Lin D, Linder ME, Ling J, Liu J, Liu J, London M, Ma H, Mao J, McDade JE, McMillan A, Moore AM, Oh WC, Ouyang Y, Patel R, Paul M, Paulsen LC, Qiu J, Rhee A, Rubashkin MG, Soh IY, Sotuyo NE, Srinivas V, Suarez A, Wong A, Wong R, Xie WR, Xu Y, Yu AT, Koszul R, Bader JS, Boeke JD, Chandrasegaran S (2014) Total synthesis of a functional designer eukaryotic chromosome. Science 344(6179):55–58
24. Li MZ, Elledge SJ (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nat Methods 4(3):251–256
25. Shao Z, Zhao H, Zhao H (2009) DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. Nucleic Acids Res 37(2):e16
26. Wingler LM, Cornish VW (2011) Reiterative recombination for the in vivo assembly of libraries of multi-

gene pathways. Proc Natl Acad Sci U S A 108(37):15135–15140

27. Koschubs T, Lorenzen K, Baumli S, Sandstrom S, Heck AJ, Cramer P (2010) Preparation and topology of the mediator middle module. Nucleic Acids Res 38(10):3186–3195
28. Janke C, Magiera MM, Rathfelder N, Taxis C, Reber S, Maekawa H, Moreno-Borchart A, Doenges G, Schwob E, Schiebel E, Knop M (2004) A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. Yeast 21(11):947–962
29. van Ooyen AJ, Dekker P, Huang M, Olsthoorn MM, Jacobs DI, Colussi PA, Taron CH (2006) Heterologous protein production in the yeast Kluyveromyces lactis. FEMS Yeast Res 6(3):381–392
30. Morlino GB, Tizzani L, Fleer R, Frontali L, Bianchi MM (1999) Inducible amplification of gene copy number and heterologous protein production in the yeast Kluyveromyces lactis. Appl Environ Microbiol 65(11):4808–4813
31. Lane MM, Morrissey JP (2010) Kluyveromyces marxianus: a yeast emerging from its sister's shadow. Fungal Biol Rev 24(1–2):17–26
32. Becerra M, Baroli B, Fadda AM, Blanco Méndez J, González Siso MI (2001) Lactose bioconversion by calcium-alginate immobilization of Kluyveromyces lactis cells. Enzym Microb Technol 29(8–9):506–512
33. Colussi PA, Taron CH (2005) Kluyveromyces lactis LAC4 promoter variants that lack function in bacteria but retain full function in K. lactis. Appl Environ Microbiol 71(11):7092–7098
34. Read JD, Colussi PA, Ganatra MB, Taron CH (2007) Acetamide selection of Kluyveromyces lactis cells transformed with an integrative vector leads to high-frequency formation of multicopy strains. Appl Environ Microbiol 73(16):5088–5096
35. Hoshida H, Murakami N, Suzuki A, Tamura R, Asakawa J, Abdel-Banat BM, Nonklang S, Nakamura M, Akada R (2014) Non-homologous end joining-mediated functional marker selection for DNA cloning in the yeast Kluyveromyces marxianus. Yeast 31(1):29–46
36. Hong J, Wang Y, Kumagai H, Tamaki H (2007) Construction of thermotolerant yeast expressing thermostable cellulase genes. J Biotechnol 130(2):114–123
37. Faraco V, Ercole C, Festa G, Giardina P, Piscitelli A, Sannia G (2008) Heterologous expression of heterodimeric laccase from Pleurotus ostreatus in Kluyveromyces lactis. Appl Microbiol Biotechnol 77(6):1329–1335
38. Falcone C, Saliola M, Chen XJ, Frontali L, Fukuhara H (1986) Analysis of a 1.6-micron circular plasmid from the yeast Kluyveromyces drosophilarum: structure and molecular dimorphism. Plasmid 15(3):248–252
39. Madzak C, Beckerich J-M (2013) Heterologous protein expression and secretion in Yarrowia lipolytica. In: Barth G (ed) Yarrowia lipolytica, vol 25, Microbiology Monographs. Springer, Berlin, pp 1–76
40. Barth G, Gaillardin C (1996) Yarrowia lipolytica. In: Wolf K (ed) Nonconventional yeasts in biotechnology. Springer, Berlin, pp 313–338
41. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wesolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet JL (2004) Genome evolution in yeasts. Nature 430(6995):35–44
42. Dominguez A, Ferminan E, Sanchez M, Gonzalez FJ, Perez-Campo FM, Garcia S, Herrero AB, San Vicente A, Cabello J, Prado M, Iglesias FJ, Choupina A, Burguillo FJ, Fernandez-Lago L, Lopez MC (1998) Non-conventional yeasts as hosts for heterologous protein production. Int Microbiol 1(2):131–142
43. Tharaud C, Ribet AM, Costes C, Gaillardin C (1992) Secretion of human blood coagulation factor XIIIa by the yeast Yarrowia lipolytica. Gene 121(1):111–119
44. Franke AE, Kaczmarek FS, Eisenhard ME, Geoghegan KF, Danley DE, DeZeeuw JR, O'Donnell MM, Gollaher MG, Davidow LS (1988) Expression and secretion of bovine prochymosin in Yarrowia lipolytica. Dev Ind Microbiol 29:43–57
45. Davidow LS, Franke AE, DeZeeuw JR (1985) New Yarrowia lipolytica transformants used for expression and secretion of heterologous proteins especially prorennin and human anaphylatoxin C5a. USPTO Patent US4937189A
46. Madzak C, Gaillardin C, Beckerich JM (2004) Heterologous protein expression and secretion in the non-conventional yeast Yarrowia lipolytica: a review. J Biotechnol 109(1–2):63–81
47. Chuang LT, Chen DC, Nicaud JM, Madzak C, Chen YH, Huang YS (2010) Co-expression of heterologous desaturase genes in Yarrowia lipolytica. New Biotechnol 27(4):277–282
48. Ahmad M, Hirz M, Pichler H, Schwab H (2014) Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production. Appl Microbiol Biotechnol 98(12):5301–5317
49. Spohner SC, Muller H, Quitmann H, Czermak P (2015) Expression of enzymes for the usage in food and feed industry with Pichia pastoris. J Biotechnol 202:118–134

50. Cos O, Ramon R, Montesinos JL, Valero F (2006) Operational strategies, monitoring and control of heterologous protein production in the methylotrophic yeast Pichia pastoris under different promoters: a review. Microb Cell Factories 5:17
51. Jacob D, Ruffie C, Dubois M, Combredet C, Amino R, Formaglio P, Gorgette O, Pehau-Arnaudet G, Guery C, Puijalon O, Barale JC, Menard R, Tangy F, Sala M (2014) Whole Pichia pastoris yeast expressing measles virus nucleoprotein as a production and delivery system to multimerize Plasmodium antigens. PLoS ONE 9(1):e86658
52. Daly R, Hearn MT (2005) Expression of heterologous proteins in Pichia pastoris: a useful experimental tool in protein engineering and production. J Mol Recognit 18(2):119–138
53. Sugiki T, Ichikawa O, Miyazawa-Onami M, Shimada I, Takahashi H (2012) Isotopic labeling of heterologous proteins in the yeast Pichia pastoris and Kluyveromyces lactis. Methods Mol Biol 831:19–36
54. Cereghino JL, Cregg JM (2000) Heterologous protein expression in the methylotrophic yeast Pichia pastoris. FEMS Microbiol Rev 24(1):45–66
55. Krainer FW, Dietzsch C, Hajek T, Herwig C, Spadiut O, Glieder A (2012) Recombinant protein expression in Pichia pastoris strains with an engineered methanol utilization pathway. Microb Cell Factories 11:22
56. Resina D, Maurer M, Cos O, Arnau C, Carnicer M, Marx H, Gasser B, Valero F, Mattanovich D, Ferrer P (2009) Engineering of bottlenecks in Rhizopus oryzae lipase production in Pichia pastoris using the nitrogen source-regulated FLD1 promoter. Nat Biotechnol 25(6):396–403
57. Shen S, Sulter G, Jeffries TW, Cregg JM (1998) A strong nitrogen source-regulated promoter for controlled expression of foreign genes in the yeast Pichia pastoris. Gene 216(1):93–102
58. Waterham HR, Digan ME, Koutz PJ, Lair SV, Cregg JM (1997) Isolation of the Pichia pastoris glyceraldehyde-3-phosphate dehydrogenase gene and regulation and use of its promoter. Gene 186(1):37–44
59. Menendez J, Valdes I, Cabrera N (2003) The ICL1 gene of Pichia pastoris, transcriptional regulation and use of its promoter. Yeast 20(13):1097–1108
60. Macauley-Patrick S, Fazenda ML, McNeil B, Harvey LM (2005) Heterologous protein production using the Pichia pastoris expression system. Yeast 22(4):249–270
61. Nett JH, Cook WJ, Chen MT, Davidson RC, Bobrowicz P, Kett W, Brevnova E, Potgieter TI, Mellon MT, Prinz B, Choi BK, Zha D, Burnina I, Bukowski JT, Du M, Wildt S, Hamilton SR (2013) Characterization of the Pichia pastoris protein-O-mannosyltransferase gene family. PLoS ONE 8(7):e68325
62. Bobrowicz P, Davidson RC, Li H, Potgieter TI, Nett JH, Hamilton SR, Stadheim TA, Miele RG, Bobrowicz B, Mitchell T, Rausch S, Renfer E, Wildt S (2004) Engineering of an artificial glycosylation pathway blocked in core oligosaccharide assembly in the yeast Pichia pastoris: production of complex humanized glycoproteins with terminal galactose. Glycobiology 14(9):757–766
63. Wu S, Letchworth GJ (2004) High efficiency transformation by electroporation of Pichia pastoris pretreated with lithium acetate and dithiothreitol. BioTech 36(1):152–154
64. Shen Q, Wu M, Wang HB, Naranmandura H, Chen SQ (2012) The effect of gene copy number and co-expression of chaperone on production of albumin fusion proteins in Pichia pastoris. Appl Microbiol Biotechnol 96(3):763–772
65. Kobayashi K, Kuwae S, Ohya T, Ohda T, Ohyama M, Ohi H, Tomomitsu K, Ohmura T (2000) High-level expression of recombinant human serum albumin from the methylotrophic yeast Pichia pastoris with minimal protease production and activation. J Biosci Bioeng 89(1):55–61
66. Scorer CA, Clare JJ, McCombie WR, Romanos MA, Sreekrishna K (1994) Rapid selection using G418 of high copy number transformants of Pichia pastoris for high-level foreign gene expression. BioTech 12(2):181–184
67. Abad S, Kitz K, Hormann A, Schreiner U, Hartner FS, Glieder A (2010) Real-time PCR-based determination of gene copy numbers in Pichia pastoris. Biotechnol J 5(4):413–420
68. Lin-Cereghino GP, Stark CM, Kim D, Chang J, Shaheen N, Poerwanto H, Agari K, Moua P, Low LK, Tran N, Huang AD, Nattestad M, Oshiro KT, Chang JW, Chavan A, Tsai JW, Lin-Cereghino J (2013) The effect of alpha-mating factor secretion signal mutations on recombinant protein expression in Pichia pastoris. Gene 519(2):311–317
69. Heimo H, Palmu K, Suominen I (1997) Expression in Pichia pastoris and purification of Aspergillus awamori glucoamylase catalytic domain. Protein Expr Purif 10(1):70–79
70. Bieszke JA, Spudich EN, Scott KL, Borkovich KA, Spudich JL (1999) A eukaryotic protein, NOP-1, binds retinal to form an archaeal rhodopsin-like photochemically reactive pigment. Biochemistry 38(43):14138–14145
71. Boze H, Laborde C, Chemardin P, Richard F, Venturin C, Combarnous Y, Moulin G (2001) High-level secretory production of recombinant porcine follicle-stimulating hormone by Pichia pastoris. Process Biochem 36:907–913
72. Baumann K, Carnicer M, Dragosits M, Graf AB, Stadlmann J, Jouhten P, Maaheimo H, Gasser B, Albiol J, Mattanovich D, Ferrer P (2010) A multilevel study of recombinant Pichia pastoris in different oxygen conditions. BMC Syst Biol 4:141
73. Aller SG, Unger VM (2006) Projection structure of the human copper transporter CTR1 at 6-A resolution

reveals a compact trimer with a novel channel-like architecture. Proc Natl Acad Sci U S A 103(10):3627–3632

74. Long SB, Tao X, Campbell EB, MacKinnon R (2007) Atomic structure of a voltage-dependent K+ channel in a lipid membrane-like environment. Nature 450(7168):376–382
75. Singh S, Zhang M, Bertheleme N, Strange PG, Byrne B (2012) Purification of the human G protein-coupled receptor adenosine A(2a)R in a stable and functional form expressed in Pichia pastoris. Curr Protoc Protein Sci. Chapter 29:Unit 29.24
76. Opekarova M, Tanner W (2003) Specific lipid requirements of membrane proteins – a putative bottleneck in heterologous expression. Biochim Biophys Acta 1610(1):11–22
77. Hirz M, Richter G, Leitner E, Wriessnegger T, Pichler H (2013) A novel cholesterol-producing Pichia pastoris strain is an ideal host for functional expression of human Na, K-ATPase alpha3beta1 isoform. Appl Microbiol Biotechnol 97(21):9465–9478
78. Zhou H, Chen Z, Chen H, Li S, Huang B, Bi R (2007) Co-expression and purification of recombinant human insulin-like growth factor II and insulin-like growth factor binding protein-6 in Pichia pastoris yeast. Protein Pept Lett 14(9):876–880
79. Jamshad M, Rajesh S, Stamataki Z, McKeating JA, Dafforn T, Overduin M, Bill RM (2008) Structural characterization of recombinant human CD81 produced in Pichia pastoris. Protein Expr Purif 57(2):206–216
80. Bonander N, Jamshad M, Oberthur D, Clare M, Barwell J, Hu K, Farquhar MJ, Stamataki Z, Harris HJ, Dierks K, Dafforn TR, Betzel C, McKeating JA, Bill RM (2013) Production, purification and characterization of recombinant, full-length human claudin-1. PLoS ONE 8(5):e64517
81. Gellissen G (2000) Heterologous protein production in methylotrophic yeasts. Appl Microbiol Biotechnol 54(6):741–750
82. Mayer AF, Hellmuth K, Schlieker H, Lopez-Ulibarri R, Oertel S, Dahlems U, Strasser AW, van Loon AP (1999) An expression system matures: a highly efficient and cost-effective process for phytase production by recombinant strains of Hansenula polymorpha. Biotechnol Bioeng 63(3):373–381
83. Amuel C, Gellissen G, Hollenberg CP, Suckow M (2000) Analysis of heat shock promoters in Hansenula polymorpha: the TPS1 promoter, a novel element for heterologous gene expression. Biotechnol Bioprocess Eng 5:247–252
84. Muller S, Sandal T, Kamp-Hansen P, Dalboge H (1998) Comparison of expression systems in the yeasts Saccharomyces cerevisiae, Hansenula polymorpha, Klyveromyces lactis. Schizosaccharomyces pombe and Yarrowia lipolytica. Cloning of two novel promoters from Yarrowia lipolytica. Yeast 14(14):1267–1283
85. Boer E, Steinborn G, Matros A, Mock HP, Gellissen G, Kunze G (2007) Production of interleukin-6 in Arxula adeninivorans, Hansenula polymorpha and Saccharomyces cerevisiae by applying the wide-range yeast vector (CoMed) system to simultaneous comparative assessment. FEMS Yeast Res 7(7):1181–1187
86. Li Y, Song H, Li J, Wang Y, Yan X, Zhao B, Zhang X, Wang S, Chen L, Qiu B, Meng S (2011) Hansenula polymorpha expressed heat shock protein gp96 exerts potent T cell activation activity as an adjuvant. J Biotechnol 151(4):343–349
87. Gellissen GMP, Dahlems U, Jenzelewski V, Gavagan JE, DiCosimo R, Anton DL, Janowicz ZA Recombinant (1996) Hansenula polymorpha as a biocatalyst: coexpression of the spinach glycolate oxidase (GO) and the S. cerevisiae catalase T (CTT1) gene. Appl Environ Microbiol 46(1):46–54
88. Klabunde J, Diesel A, Waschk D, Gellissen G, Hollenberg CP, Suckow M (2002) Single-step co-integration of multiple expressible heterologous genes into the ribosomal DNA of the methylotrophic yeast Hansenula polymorpha. Appl Microbiol Biotechnol 58(6):797–805
89. Gellisen G et al (2005) FEMS Yeast Res 5:1079–1096, Elsevier

10 *Leishmania tarentolae* for the Production of Multi-subunit Complexes

Tomoaki Niimi

Abstract

Multi-subunit protein complexes are involved in a wide variety of cellular processes including DNA replication, transcriptional regulation, signal transduction, protein folding and degradation. A better understanding of the function of these protein complexes requires structural insights into the molecular arrangement and interactions of their constituent subunits. However, biochemical and structural analysis of multi-subunit protein complexes is still limited because of technical difficulties with their recombinant expression and reconstitution. This chapter presents an overview of a novel protein expression system based on *Leishmania tarentolae*, a unicellular protozoan parasite of lizards, and practical considerations for the production of multi-subunit protein complexes. The *Leishmania tarentolae* expression system offers fully eukaryotic protein expression with post-translational modifications but with ease of handling similar to bacteria. This chapter also summarizes studies on the production of laminins, large heterotrimeric glycoproteins of the extracellular matrix, using this expression system. In addition, a recently developed *Leishmania tarentolae*-based cell-free translation system is briefly described.

Keywords

Cell-free translation • Heterologous protein expression • Laminin chain assembly • Protozoan parasite • Trypanosomatids

10.1 Introduction

Protozoan parasites of the genus *Leishmania* belong to the family Trypanosomatidae, which comprises unicellular organisms characterized by the presence of a single flagellum and a unique mitochondrial DNA-containing organelle, the

T. Niimi (✉)
Graduate School of Bioagricultural Sciences,
Nagoya University, Nagoya, Japan
e-mail: tniimi@agr.nagoya-u.ac.jp

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_10

kinetoplast [1–4]. *Leishmania* parasites are transmitted to their vertebrate hosts by the bite of infected female phlebotomine sandflies, and alternates between two life-forms: an extracellular promastigote form living in the digestive tract of female sandflies and an intracellular amastigote form residing in vertebrate macrophages. More than 20 *Leishmania* species have been reported to cause a wide spectrum of tropical diseases, collectively termed leishmaniasis [5–7]. According to a recent report from the World Health Organization, there are an estimated 1.3 million new cases of, and 20,000–30,000 deaths from leishmaniasis annually in 98 countries worldwide [8].

Not all members of genus *Leishmania* are parasites of mammals. *Leishmania tarentolae* (*L. tarentolae*) is a lizard-infecting species, which was first isolated from the Moorish gecko, *Tarentola mauritanica* (Fig. 10.1) [9]. *L. tarentolae* does not cause pathology in humans nor in severe combined immunodeficient (SCID) mice [10]. This lack of pathogenicity has been addressed in several studies comparing *L. tarentolae* with pathogenic *Leishmania* species [11–13]. One study showed that several of the genes expressed preferentially in the intracellular amastigote form of pathogenic *Leishmania* species are lacking from *L. tarentolae*, providing a possible explanation for why *L. tarentolae* is unable to replicate efficiently in mammalian macrophages [13].

As *L. tarentolae* grows rapidly in simple nutrient media and is not pathogenic to humans, it has been used as a model organism for studying unique features of Trypanosomatids, such as RNA editing and polycistronic transcription [14–16]. In Trypanosomatids, messenger RNAs (mRNAs) are transcribed as polycistronic precursors that are post-transcriptionally processed into individual mRNAs by trans-splicing and polyadenylation (Fig. 10.2) [17, 18]. Trans-splicing adds a capped 39-nucleotide spliced leader sequence to the 5′ end of the mRNA, which is necessary for RNA transport, stability, and translation efficiency. In these organisms, regulation of gene expression occurs predominantly post-transcriptionally through the structure of the intergenic untranslated regions (UTRs) [19].

Using the unique features of protozoan parasites, Breitling et al. developed a novel eukaryotic expression system based on *L. tarentolae* for the production of recombinant proteins [20]. Constitutive or inducible expression of target proteins for the cytosolic or secretory pathway is possible, and the expression vector can be either stably integrated into the genome or maintained episomally [21, 22]. Now that the *L. tarentolae* protein expression system has been commercialized by Jena Bioscience GmbH (http://www.jenabioscience.com/), it is being used more widely. There are increasing reports of heterologous protein expression in *L. tarentolae*; however, few examples of the expression of multi-subunit

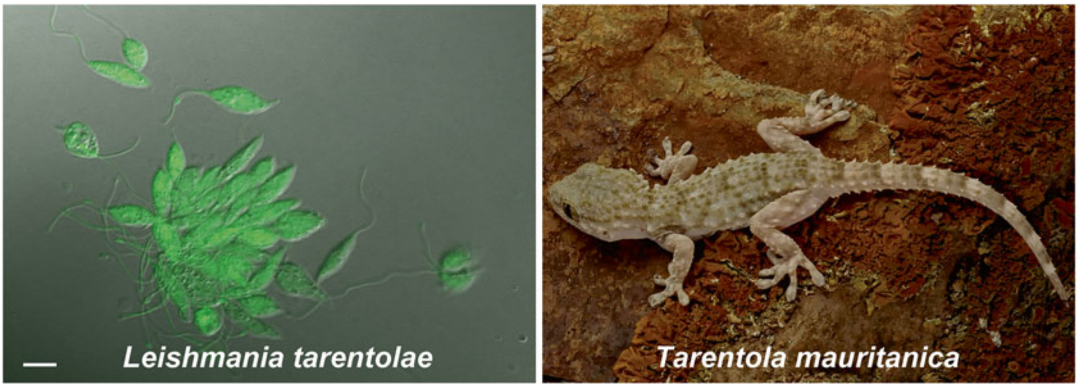


Fig. 10.1 ***L. tarentolae* cells expressing green fluorescent protein (GFP) and their natural host, *Tarentola mauritanica* .** Scale bar = 5 μm

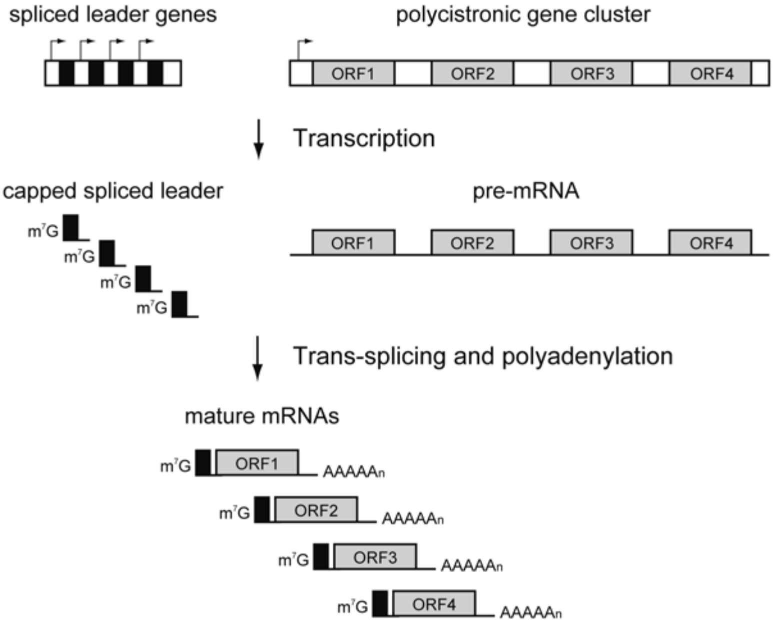


Fig. 10.2 The general process of mRNA synthesis in Trypanosomatids. Most *Leishmania* genes are organized in tandem repeats as indicated by the *grey boxes*. At another chromosomal location there are several hundred tandem direct repeats encoding spliced leader RNAs (*black boxes*). In some cases, a single promoter is present upstream of the first gene in the cluster, whereas in the spliced leader gene cluster each repeat is presumed to bear a transcriptional promoter. After transcription, the polycistronic pre-mRNA is processed by trans-splicing and polyadenylation to produce mature mRNAs

proteins have been reported [23–30]. This chapter outlines the *L. tarentolae* expression system and our research on multi-subunit protein expression in this system.

10.2 Maintenance of *L. tarentolae* Cells

The culture of *L. tarentolae* is much easier than that of mammalian cells. Because it is not pathogenic to mammals, it requires only biosafety level 1 facilities. *L. tarentolae* cells can be grown in brain–heart infusion broth without serum but supplemented with hemin, which is essential for growth of heme-deficient organisms such as Trypanosomatids [31, 32]. *L. tarentolae* cells require aerobic conditions, so the cells are maintained in a suspension culture in ventilated tissue culture flasks. A CO_2 incubator is not necessary. Conventional static cultures are incubated in the dark at 26 °C. Agitated cultures for protein expression are incubated on an orbital shaker at 140 rpm using Erlenmeyer flasks. *L. tarentolae* cells can be grown indefinitely *in vitro*, with a doubling time of around 5 h and to high cell densities in shaking culture (approximately 5×10^8 cells/mL).

10.3 Expression Vectors for the *L. tarentolae* Expression System

In most eukaryotes, ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are transcribed by RNA polymerase I and III, respectively, and the protein-coding genes are transcribed by RNA polymerase II to yield mRNAs [33]. In Trypanosomatids, however, translation of RNA polymerase I-generated transcripts is possible because of trans-splicing of polycistronic pre-mRNAs [17]. In a *L. tarentolae* expression system, integration of an expression cassette into the chromosomal small-subunit (ssu) rRNA locus enables the generation of large numbers of transcripts for constitutive expression of target proteins [20]. Thus, the expression cassette is flanked by two fragments of the ssu rRNA locus for homologous recombination, and contains three optimized UTRs, flanking the target and marker gene insertion sites, which provide the trans-splicing signal (Fig. 10.3).

In this system, alternative cloning strategies allow heterologous proteins to be expressed cytosolically or secreted into the medium. To promote secretion, a signal sequence derived from the secreted acid phosphatase of *L. mexicana* is uti-

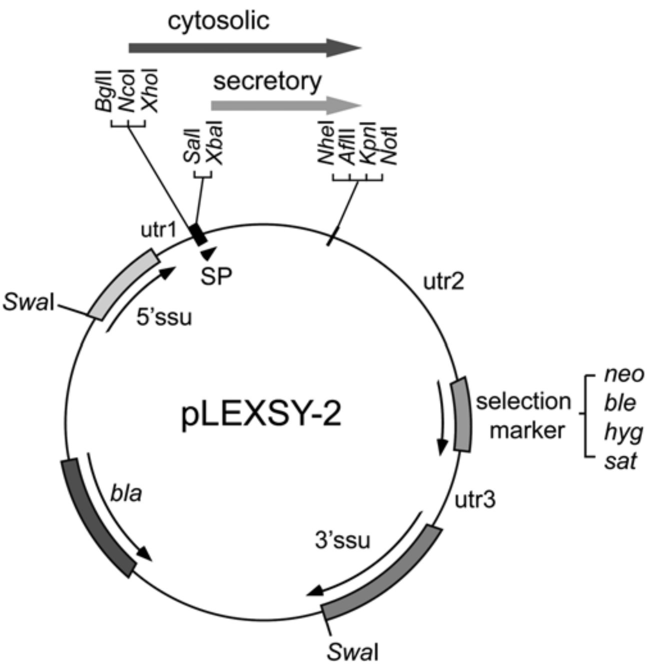


Fig. 10.3 **Map of the *L. tarentolae* expression vector.** 5′ssu and 3′ssu are regions for homologous recombination into the host chromosome following linearization of the expression plasmid with *Swa*I. The vector contains three optimized intergenic untranslated regions (UTRs) for post-transcriptional mRNA processing: utr1 derived from the 0.4 kb intergenic region (IR) of the *L. tarentolae* adenine phosphoribosyl transferase gene; utr2 from the 1.4 kb IR of the calmodulin cluster containing three tandemly arranged calmodulin genes; and utr3 from the 1.7 kb IR of the *L. major* dihydrofolate reductase–thymidylate synthase gene. SP indicates the signal peptide of *L. mexicana* secreted acid phosphatase. Alternative cloning strategies result in cytosolic or secretory expression of the target protein. As selection markers, four genes are available: *neo*, *ble*, *hyg*, and *sat* genes

lized [34]; however, the native signal sequence has also been used successfully for secretion of several proteins [20, 35]. Use of the native signal sequence may enable native processing of proteins at the N terminus.

After construction, the expression plasmid is linearized and integrated into the genome of *L. tarentolae* by homologous recombination. *L. tarentolae* cells can be routinely transfected with plasmid DNA by electroporation and the transfected cells can be selected with antibiotics. Currently, four selectable marker genes are available in this system: neomycin phosphotransferase (*neo*), hygromycin phosphotransferase (*hyg*), bleomycin resistance protein (*ble*), and streptothricin acetyltransferase (*sat*) that confer resistance to G418, hygromycin, bleomycin, and nourseothricin, respectively. Therefore, up to four genes can be simultaneously expressed to produce multi-subunit proteins. If additional selection marker genes, for example, puromycin acetyltransferase (*pac*), are incorporated into the expression vector, more than four genes could be simultaneously expressed. The construction of markerless *L. tarentolae* strains carrying multiple expression cassettes may be possible, but is known to be difficult [36]. Platforms for inducible expression of target proteins are available; however, only one, or at most two, selectable marker genes are offered.

10.4 Post-translational Modifications (PTMs) in *L. tarentolae*

In contrast to prokaryotic systems, expression of recombinant proteins in eukaryotic expression systems allows PTMs. Protein glycosylation is one of the most common PTMs in eukaryotes, and it plays essential roles in many biological processes, such as cell recognition, cell-cell communication, signaling, embryo development, and immunity [37, 38]. The pattern of N-linked glycosylation of glycoproteins is important because the number and position of N-linked oligosaccharides often have significant effects on protein function [39, 40]. *L. tarentolae* has been reported to produce higher eukaryote-like biantennary N-glycans with terminal galactose and core fucose but lacking sialic acid residues, indicating that the N-glycosylation pathway of *L. tarentolae* is more similar to that in mammals than in yeast and insect cells [20]. It has been recently demonstrated that human soluble amyloid precursor protein α (sAPPα) produced in *L. tarentolae* was both N- and O-glycosylated on similar sites as described for mammalian-expressed sAPPα [41]. However, more complex O-glycan structures commonly found in mammalian expression systems were not observed. This insufficient glycosylation is probably due to the lack of certain glycosyltransferase activities in *L. tarentolae* [20]. Genetic engineering of the *L. tarentolae* host or *in vitro* glycosylation using specific glycosyltransferases may provide a method for producing glycoproteins with more complex glycan structures.

L. tarentolae has the potential to perform other eukaryotic PTMs, including processing of signal sequences, proper protein folding, and disulfide bond formation. Previously, we successfully produced the disulfide-linked heterotrimeric glycoprotein, laminin-332, in the *L. tarentolae* expression system [42]. In the following section, the production of multi-subunit proteins in the *L. tarentolae* expression system is described.

10.5 Production of Recombinant Laminin-332 Using the *L. tarentolae* Expression System

Laminins are a family of extracellular matrix glycoproteins localized in the basement membrane, and bind to cell surface receptors such as integrins, supporting various cellular functions including adhesion, migration, proliferation, and differentiation [43–45]. They consist of three subunits, α, β, and γ chains, which bind to each other via disulfide bonds between laminin coiled-coil (LCC) domains to form a cross-shaped structure with three short arms and one long arm (Fig. 10.4a). To date, five α, three β, and three γ chains have been identified to combine into at least 16 heterotrimeric isoforms [46, 47]. They are named according to their chain composition; for example, laminin-111 consists of α1, β1, and γ1 chains.

There is a great need to develop a method for the efficient and mass production of recombinant laminins because some laminin isoforms (laminin-511 and -332) are able to support the stable culture of human embryonic stem cells (hESCs) and human induced pluripotent stem cells (hiPSCs) [48–51]. Chemically defined, xeno-free, and feeder-free culture systems are required for future use of hESCs and hiPSCs in regenerative medicine, and recombinant laminins can be used to replace feeder cells. Because they are large heterotrimeric proteins (400–900 kDa), it is difficult to express correctly folded laminins in bacterial and yeast expression systems. Thus, mammalian cells have been used to prepare recombinant laminins; however, mass production of recombinant laminins remains laborious.

The *L. tarentolae* expression system combines the ease of handling found with bacteria and yeast, with eukaryotic protein folding and mammalian-type PTMs of target proteins. These advantages of the *L. tarentolae* system prompted us to examine whether recombinant laminins produced in *L. tarentolae* acquire proper conformation and bioactivity. As a model for the production of laminins in *L. tarentolae*, laminin-332, which consists of α3, β3, and γ2 chains, was selected because it is the smallest laminin isoform, with truncated short arms in all three chains (Fig. 10.4a). The full-length cDNAs of laminin β3 and γ2 chains, without signal sequences, were cloned into the *L. tarentolae* expression vector (Fig. 10.3) behind the signal sequence of *L. mexicana* secreted acid phosphatase. The α3 chain undergoes extracellular proteolytic processing of both ends in mammals [52–54], so the cDNA of the α3 chain containing the fully processed form was cloned into the expression vector in the same way as β3 and γ2 chains. These three plasmids were sequentially transfected into *L. tarentolae* cells by electroporation, and stable transfectants were selected by culturing cells on solid media with three antibiotics.

For the assembly of the laminin chains, the β and γ chains first assemble to form heterodimers in the endoplasmic reticulum after translation of individual chains [55]. Subsequently, one α chain joins the β–γ heterodimers to form α–β–γ heterotrimers, which are transported through the

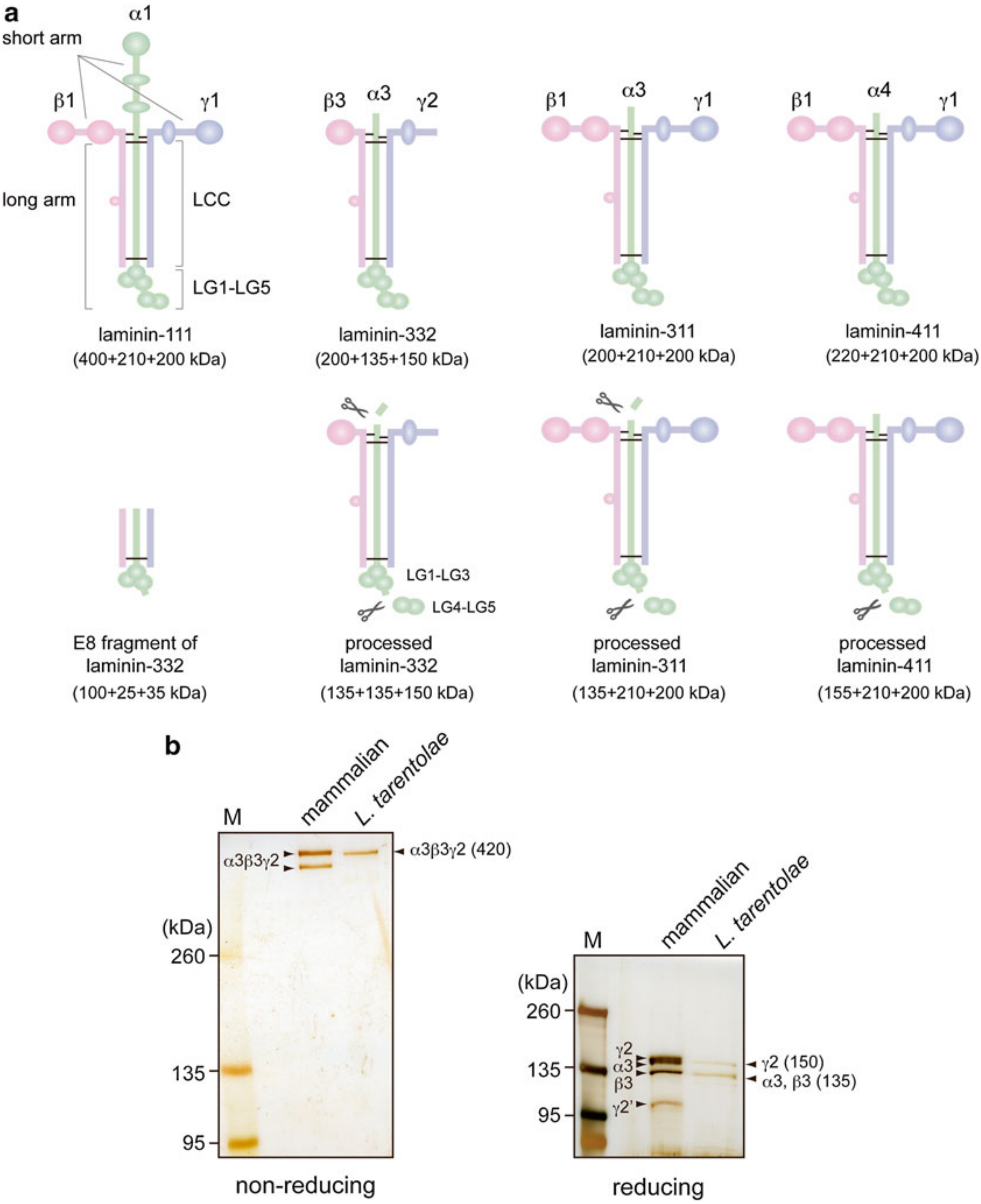


Fig. 10.4 Production of recombinant laminins using the *L. tarentolae* expression system. (**a**) Schematic structure of laminin isoforms. The α, β, and γ chains assemble to form a triple-stranded α-helical coiled-coil structure in the laminin coiled-coil (LCC) domain. The laminin globular (LG) domains are typically involved in cellular interactions. The size of each chain is shown below the laminins. Physiological cleavage by enzymes known to occur for the α3 and α4 chains is indicated by scissors. The laminin isoforms produced successfully in *L. tarentolae* expression system are shown in the lower panel. (**b**) Purified laminin-332 from mammalian 293-F cells (*left lane*) and *L. tarentolae* cells (*right lane*) were separated by SDS-PAGE under non-reducing or reducing conditions and analyzed by silver staining

secretory pathway. Single cysteine residues at the C termini of the β and γ chains form an interchain disulfide bond. At the N termini of the LCC domains, all three chains have two cysteine residues and are disulfide linked to each other before secretion. Accordingly, the heterotrimer can be viewed by SDS-PAGE under reducing and nonreducing conditions. The recombinant *L. tarentolae* strain, harboring the three subunits of laminin-332, efficiently formed α3–β3–γ2 heterotrimers (~420 kDa) with disulfide bonds and secreted it into the medium, demonstrating for the first time that the three chains of semi-intact laminin can form heterotrimers in a unicellular eukaryote (Fig. 10.4b) [42]. Hydrophobic interactions within the α-helical coiled-coils are the main driving force for laminin chain assembly, so synthetic peptides or small fragments of the LCC domains of the three subunits can assemble themselves *in vitro* [55]. However, assembly of the whole unprocessed laminin chains is difficult to achieve *in vitro* or in bacterial and yeast expression systems, probably because the individual chains need to fold correctly before assembly. Correct folding is also required to facilitate the proper positioning of cysteine residues, which allows the correct formation of intra-chain disulfide bonds in the short arm region of all three chains and in the laminin globular (LG) domain at the N-terminal end of the α chain. As we were able to efficiently form laminin heterotrimers, *L. tarentolae* cells may provide the appropriate molecular chaperones to aid proper protein folding as well as a transport system for large proteins. When analyzing the recombinant *L. tarentolae* strains harboring only β3 or β3/γ2 subunits, the β3 monomers and the β3–γ2 heterodimers were detected in the cells but not secreted into the medium, suggesting that the monitoring system that allows only heterotrimers to be transported through the secretory pathway is also present in *L. tarentolae* cells [42]. The purified laminin-332 showed similar cell adhesion activity to laminin-332 purified from mammalian cells [42]. The production yield (about 0.5 mg per liter of culture medium) was also similar to that of mammalian cells [42].

10.6 Production of Other Laminins Using the *L. tarentolae* Expression System

The successful production of laminin-332 led us to investigate whether other isoforms of the laminin family could be produced in *L. tarentolae* (Fig. 10.4a). When the β1 and γ1 chains were swapped with the β3 and γ2 chains in the recombinant *L. tarentolae* strain, laminin-311 (~545 kDa) was formed and secreted into the culture medium [42]. In addition, α4 chain without the LG4–LG5 domains could assemble with β1–γ1 heterodimers to form laminin-411 (~565 kDa) (unpublished observation). However, it was difficult to express other α chains with or without LG4–LG5 domains in *L. tarentolae.* In fact, intact α chains can be often expressed but not folded correctly, and then are unable to assemble with β–γ heterodimers. These results suggest that larger and more complex laminin chains could not be expressed in *L. tarentolae* cells. Thus far, laminin-411 (~565 kDa) is the largest recombinant protein with multiple subunits produced in the *L. tarentolae* expression system.

The expression level of laminin-332 in *L. tarentolae* was insufficient for mass production despite it being the smallest laminin. Therefore, the laminin E8 fragment, which is a truncated laminin composed of the C-terminal regions of all three chains, was expressed in *L. tarentolae* (Fig. 10.4a). It contains the active integrin-binding site but lacks other activities of whole laminins such as heparin-binding activity; therefore, it serves as a functionally minimal form that efficiently maintains hESCs and hiPSCs [50, 51]. When the three chains corresponding to the E8 fragment of laminin-332 were expressed in *L. tarentolae*, these chains successfully formed heterotrimers (~160 kDa) and were secreted into the culture medium (unpublished observation). The expression level was at most twice that of the processed form of laminin-332. For high-level expression of laminin heterotrimers in the *L. tarentolae* expression system, careful optimization of culture conditions might be required.

10.7 An *L. tarentolae*-Based Cell-Free Translation System

Cell-free translation systems offer several advantages over cell-based expression systems, including the synthesis of difficult targets, such as toxins and membrane proteins, the easy modification of reaction conditions, suitability for high-throughput strategies, and rapid production [56–58]. Although any organisms could potentially be used as sources for the preparation of a cell-free translation systems, the most popular are those based on *Escherichia coli*, wheat germ, and rabbit reticulocytes. The choice of the system should be determined by the biochemical nature of the target protein and the downstream application. In general, *Escherichia coli*-based systems provide higher yields and more homogeneous proteins suitable for structural studies. Eukaryotic cell-free systems, although less productive, provide a better platform for functional studies, particularly for proteins with PTMs. We have previously expressed laminin-332 subunits in a cell-free translation system based on insect cell extract [59]. β3–γ2 LCC domain heterodimers (~130 kDa) and α3–β3–γ2 LCC domain heterotrimers (~200 kDa) were successfully formed with disulfide bonds following co-translation of each chain, however, intact β3 and γ2 chains were unable to form β3–γ2 heterodimers, indicating that the proper folding of laminin-332 subunits that included the short arm region was deficient in this system.

Recently, Alexandrov's group developed a eukaryotic cell-free translation system based on extracts of *L. tarentolae* cells [60–62]. They discovered species-independent translational sequences that mediate efficient cell-free protein synthesis in any prokaryotic and eukaryotic systems, and applied them to express proteins in *L. tarentolae* cell extract. Moreover, addition of an anti-spliced leader oligonucleotide to *L. tarentolae* cell extract suppressed the translation of endogenous *L. tarentolae* mRNAs. Using this system, Guo et al. could produce *in vitro* all six subunits of the 600 kDa HOPS and CORVET multi-subunit membrane tethering complexes [63]. This cell-free translation system is also available from Jena Bioscience GmbH. Although a limited number of proteins have been tested because of the recent development of this product, this system may be suitable for high-throughput analysis of expression of multi-subunit proteins.

10.8 Conclusions

In this chapter, the production of multi-subunit proteins using the *L. tarentolae* expression system was discussed. Laminin-332, a large heterotrimeric glycoprotein, could be produced using the *L. tarentolae* expression system, however, it was not in intact form but in processed form, suggesting that *L. tarentolae* cells do not have the same protein folding machinery as mammalian cells for expression of large proteins with complex structures like laminin α chains. Using the *L. tarentolae* expression system, laminin-332 subunits could assemble to form heterotrimers with disulfide bonds and were secreted into the culture medium, whereas it is difficult in bacterial and yeast expression systems. Thus, the *L. tarentolae* system provides an alternative platform to mammalian cells for the production of multi-subunit proteins. Up to four genes can be introduced into *L. tarentolae* cells to produce a stable cell line that can be scaled up to larger volumes. The drawbacks of this system include the limited number of expressible genes and the long experimental time line. One round of transfection and clonal selection can take up to 2 weeks (Fig. 10.5). Cell-free translation systems, where an unlimited number of genes can be co-expressed simultaneously, overcome these drawbacks. Although cell-free translation systems are relatively high cost and low yield, a recently developed cell-free system based on *L. tarentolae* enabled rapid production and reconstitution of six subunits of the multimeric membrane tethering complexes. With the range of expression systems now available, it is important for researchers to understand their advantages and disadvantages so the optimal expression systems can be selected, depending on the purpose of the target proteins [64, 65]. Both expression systems based on *L. tarentolae*

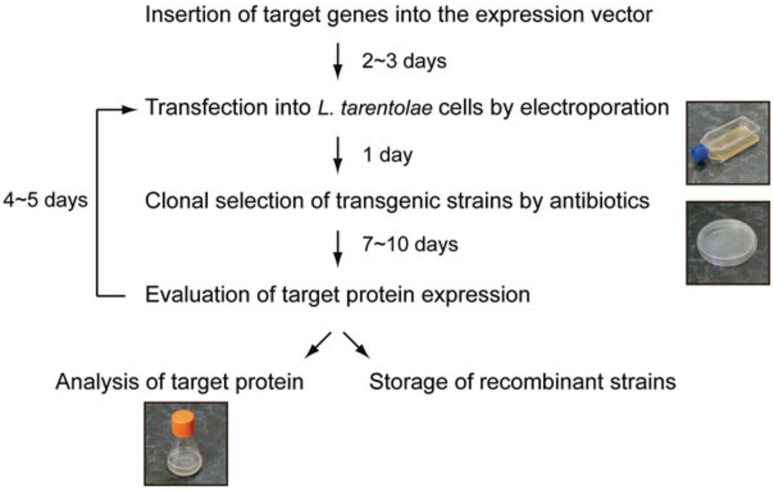

Insertion of target genes into the expression vector
2~3 days
Transfection into *L. tarentolae* cells by electroporation
1 day
4~5 days
Clonal selection of transgenic strains by antibiotics
7~10 days
Evaluation of target protein expression
Analysis of target protein
Storage of recombinant strains

Fig. 10.5 The workflow for the production of multi-subunit proteins using the *L. tarentolae* expression system

are relatively new, so there are few examples of multi-subunit protein expression using them. The structural and biochemical analysis of many other multi-subunit proteins will benefit from the use of these expression systems, and will lead to future improvements in the technology.

Acknowledgments The preparation of this chapter was supported by JSPS KAKENHI Grant Numbers 18108003 and 26350959.

References

1. Shapiro TA, Englund PT (1995) The structure and replication of kinetoplast DNA. Annu Rev Microbiol 49:117–143
2. Clayton CE (1999) Genetic manipulation of kinetoplastida. Parasitol Today 15(9):372–378
3. Beverley SM (2003) Protozomics: trypanosomatid parasite genetics comes of age. Nat Rev Genet 4(1):11–19
4. Liu B, Liu Y, Motyka SA, Agbo EE, Englund PT (2005) Fellowship of the rings: the replication of kinetoplast DNA. Trends Parasitol 21(8):363–369
5. Lipoldova M, Demant P (2006) Genetic susceptibility to infectious disease: lessons from mouse models of leishmaniasis. Nat Rev Genet 7(4):294–305
6. Banuls AL, Hide M, Prugnolle F (2007) Leishmania and the leishmaniases: a parasite genetic update and advances in taxonomy, epidemiology and pathogenicity in humans. Adv Parasitol 64:1–109
7. Kaye P, Scott P (2011) Leishmaniasis: complexity at the host-pathogen interface. Nat Rev Microbiol 9(8):604–615
8. WHO (2013) Sustaining the drive to overcome the global impact of neglected tropical diseases. Second WHO report on neglected tropical diseases
9. Elwasila M (1988) Leishmania tarentolae Wenyon, 1921 from the gecko Tarentola annularis in the Sudan. Parasitol Res 74(6):591–592
10. Breton M, Tremblay MJ, Ouellette M, Papadopoulou B (2005) Live nonpathogenic parasitic vector as a candidate vaccine against visceral leishmaniasis. Infect Immun 73(10):6372–6382
11. Tamar S, Dumas C, Papadopoulou B (2000) Chromosome structure and sequence organization between pathogenic and non-pathogenic Leishmania spp. Mol Biochem Parasitol 111(2):401–414
12. Mizbani A, Taslimi Y, Zahedifard F, Taheri T, Rafati S (2011) Effect of A2 gene on infectivity of the non-pathogenic parasite Leishmania tarentolae. Parasitol Res 109(3):793–799
13. Raymond F, Boisvert S, Roy G, Ritt JF, Legare D, Isnard A, Stanke M, Olivier M, Tremblay MJ, Papadopoulou B, Ouellette M, Corbeil J (2012) Genome sequencing of the lizard parasite Leishmania tarentolae reveals loss of genes associated to the intracellular stage of human pathogenic species. Nucleic Acids Res 40(3):1131–1147
14. Teixeira SM (1998) Control of gene expression in Trypanosomatidae. Braz J Med Biol Res 31(12):1503–1516
15. Clayton CE (2002) Life without transcriptional control? From fly to man and back again. EMBO J 21(8):1881–1888
16. Martinez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martinez LE, Manning-Cela RG, Figueroa-Angulo EE (2010) Gene expression in trypanosomatid parasites. J Biomed Biotechnol 2010:525241
17. Lee MG, Van der Ploeg LH (1997) Transcription of protein-coding genes in trypanosomes by RNA polymerase I. Annu Rev Microbiol 51:463–489

18. Teixeira SM, de Paiva RM, Kangussu-Marcolino MM, Darocha WD (2012) Trypanosomatid comparative genomics: contributions to the study of parasite biology and different parasitic diseases. Genet Mol Biol 35(1):1–17
19. Clayton C, Shapira M (2007) Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. Mol Biochem Parasitol 156(2):93–101
20. Breitling R, Klingner S, Callewaert N, Pietrucha R, Geyer A, Ehrlich G, Hartung R, Muller A, Contreras R, Beverley SM, Alexandrov K (2002) Non-pathogenic trypanosomatid protozoa as a platform for protein research and production. Protein Expr Purif 25(2):209–218
21. Kushnir S, Gase K, Breitling R, Alexandrov K (2005) Development of an inducible protein expression system based on the protozoan host Leishmania tarentolae. Protein Expr Purif 42(1):37–46
22. Kushnir S, Cirstea IC, Basiliya L, Lupilova N, Breitling R, Alexandrov K (2011) Artificial linear episome-based protein expression system for protozoon Leishmania tarentolae. Mol Biochem Parasitol 176(2):69–79
23. Soleimani M, Mahboudi F, Davoudi N, Amanzadeh A, Azizi M, Adeli A, Rastegar H, Barkhordari F, Mohajer-Maghari B (2007) Expression of human tissue plasminogen activator in the trypanosomatid protozoan Leishmania tarentolae. Biotechnol Appl Biochem 48(Pt 1):55–61
24. Ben-Abdallah M, Bondet V, Fauchereau F, Beguin P, Goubran-Botros H, Pagan C, Bourgeron T, Bellalou J (2011) Production of soluble, active acetyl serotonin methyl transferase in Leishmania tarentolae. Protein Expr Purif 75:114–118
25. Gazdag EM, Cirstea IC, Breitling R, Lukes J, Blankenfeldt W, Alexandrov K (2010) Purification and crystallization of human Cu/Zn superoxide dismutase recombinantly produced in the protozoan Leishmania tarentolae. Acta Crystallogr Sect F: Struct Biol Cryst Commun 66(Pt 8):871–877
26. Dadashipour M, Fukuta Y, Asano Y (2011) Comparative expression of wild-type and highly soluble mutant His103Leu of hydroxynitrile lyase from Manihot esculenta in prokaryotic and eukaryotic expression systems. Protein Expr Purif 77(1):92–97
27. Dortay H, Schmockel SM, Fettke J, Mueller-Roeber B (2011) Expression of human c-reactive protein in different systems and its purification from Leishmania tarentolae. Protein Expr Purif 78(1):55–60
28. Nazari R, Davoudi N (2011) Cloning and expression of truncated form of tissue plasminogen activator in Leishmania tarentolae. Biotechnol Lett 33(3):503–508
29. Baechlein C, Meemken D, Pezzoni G, Engemann C, Grummer B (2013) Expression of a truncated hepatitis E virus capsid protein in the protozoan organism Leishmania tarentolae and its application in a serological assay. J Virol Methods 193(1):238–243
30. Jorgensen ML, Friis NA, Just J, Madsen P, Petersen SV, Kristensen P (2014) Expression of single-chain variable fragments fused with the Fc-region of rabbit IgG in Leishmania tarentolae. Microb Cell Fact 13:9
31. Chang CS, Chang KP (1985) Heme requirement and acquisition by extracellular and intracellular stages of Leishmania mexicana amazonensis. Mol Biochem Parasitol 16(3):267–276
32. Fritsche C, Sitz M, Weiland N, Breitling R, Pohl HD (2007) Characterization of the growth behavior of Leishmania tarentolae: a new expression system for recombinant proteins. J Basic Microbiol 47(5):384–393
33. Kornberg RD (2007) The molecular basis of eukaryotic transcription. Proc Natl Acad Sci U S A 104(32):12955–12961
34. Wiese M, Ilg T, Lottspeich F, Overath P (1995) Ser/Thr-rich repetitive motifs as targets for phosphoglycan modifications in Leishmania mexicana secreted acid phosphatase. EMBO J 14(6):1067–1074
35. Basak A, Shervani NJ, Mbikay M, Kolajova M (2008) Recombinant proprotein convertase 4 (PC4) from Leishmania tarentolae expression system: purification, biochemical study and inhibitor design. Protein Expr Purif 60(2):117–126
36. Mureev S, Kushnir S, Kolesnikov AA, Breitling R, Alexandrov K (2007) Construction and analysis of Leishmania tarentolae transgenic strains free of selection markers. Mol Biochem Parasitol 155(2):71–83
37. Varki A (1993) Biological roles of oligosaccharides: all of the theories are correct. Glycobiology 3(2):97–130
38. Varki A (2007) Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. Nature 446(7139):1023–1029
39. Elbein AD (1991) The role of N-linked oligosaccharides in glycoprotein function. Trends Biotechnol 9(10):346–352
40. Helenius A, Aebi M (2004) Roles of N-linked glycans in the endoplasmic reticulum. Annu Rev Biochem 73:1019–1049
41. Klatt S, Rohe M, Alagesan K, Kolarich D, Konthur Z, Hartl D (2013) Production of glycosylated soluble amyloid precursor protein alpha (sAPPalpha) in Leishmania tarentolae. J Proteome Res 12(1):396–403
42. Phan HP, Sugino M, Niimi T (2009) The production of recombinant human laminin-332 in a Leishmania tarentolae expression system. Protein Expr Purif 68(1):79–84
43. Timpl R (1996) Macromolecular organization of basement membranes. Curr Opin Cell Biol 8(5):618–624
44. Miner JH, Yurchenco PD (2004) Laminin functions in tissue morphogenesis. Annu Rev Cell Dev Biol 20:255–284
45. Domogatskaya A, Rodin S, Tryggvason K (2012) Functional diversity of laminins. Annu Rev Cell Dev Biol 28:523–553
46. Aumailley M, Bruckner-Tuderman L, Carter WG, Deutzmann R, Edgar D, Ekblom P, Engel J, Engvall E, Hohenester E, Jones JC, Kleinman HK,

Marinkovich MP, Martin GR, Mayer U, Meneguzzi G, Miner JH, Miyazaki K, Patarroyo M, Paulsson M, Quaranta V, Sanes JR, Sasaki T, Sekiguchi K, Sorokin LM, Talts JF, Tryggvason K, Uitto J, Virtanen I, von der Mark K, Wewer UM, Yamada Y, Yurchenco PD (2005) A simplified laminin nomenclature. Matrix Biol 24(5):326–332

47. Aumailley M (2013) The laminin family. Cell Adh Migr 7(1):48–55
48. Miyazaki T, Futaki S, Hasegawa K, Kawasaki M, Sanzen N, Hayashi M, Kawase E, Sekiguchi K, Nakatsuji N, Suemori H (2008) Recombinant human laminin isoforms can support the undifferentiated growth of human embryonic stem cells. Biochem Biophys Res Commun 375(1):27–32
49. Rodin S, Domogatskaya A, Strom S, Hansson EM, Chien KR, Inzunza J, Hovatta O, Tryggvason K (2010) Long-term self-renewal of human pluripotent stem cells on human recombinant laminin-511. Nat Biotechnol 28(6):611–615
50. Miyazaki T, Futaki S, Suemori H, Taniguchi Y, Yamada M, Kawasaki M, Hayashi M, Kumagai H, Nakatsuji N, Sekiguchi K, Kawase E (2012) Laminin E8 fragments support efficient adhesion and expansion of dissociated human pluripotent stem cells. Nat Commun 3:1236
51. Nakagawa M, Taniguchi Y, Senda S, Takizawa N, Ichisaka T, Asano K, Morizane A, Doi D, Takahashi J, Nishizawa M, Yoshida Y, Toyoda T, Osafune K, Sekiguchi K, Yamanaka S (2014) A novel efficient feeder-free culture system for the derivation of human induced pluripotent stem cells. Sci Rep 4:3594
52. Tsubota Y, Mizushima H, Hirosaki T, Higashi S, Yasumitsu H, Miyazaki K (2000) Isolation and activity of proteolytic fragment of laminin-5 alpha3 chain. Biochem Biophys Res Commun 278(3):614–620
53. Kariya Y, Yasuda C, Nakashima Y, Ishida K, Tsubota Y, Miyazaki K (2004) Characterization of laminin 5B and NH2-terminal proteolytic fragment of its alpha3B chain: promotion of cellular adhesion, migration, and proliferation. J Biol Chem 279(23):24774–24784
54. Marinkovich MP (2007) Tumour microenvironment: laminin 332 in squamous-cell carcinoma. Nat Rev Cancer 7(5):370–380
55. Beck K, Hunter I, Engel J (1990) Structure and function of laminin: anatomy of a multidomain glycoprotein. FASEB J 4(2):148–160
56. Bernhard F, Tozawa Y (2013) Cell-free expression-making a mark. Curr Opin Struct Biol 23(3):374–380
57. Rosenblum G, Cooperman BS (2014) Engine out of the chassis: cell-free protein synthesis and its uses. FEBS Lett 588(2):261–268
58. Harbers M (2014) Wheat germ systems for cell-free protein expression. FEBS Lett 588:2762–2773
59. Phan HP, Ezure T, Ito M, Kadowaki T, Kitagawa Y, Niimi T (2008) Expression and chain assembly of human laminin-332 in an insect cell-free translation system. Biosci Biotechnol Biochem 72(7):1847–1852
60. Mureev S, Kovtun O, Nguyen UT, Alexandrov K (2009) Species-independent translational leaders facilitate cell-free expression. Nat Biotechnol 27(8):747–752
61. Kovtun O, Mureev S, Johnston W, Alexandrov K (2010) Towards the construction of expressed proteomes using a Leishmania tarentolae based cell-free expression system. PLoS One 5(12), e14388
62. Kovtun O, Mureev S, Jung W, Kubala MH, Johnston W, Alexandrov K (2011) Leishmania cell-free protein expression system. Methods 55(1):58–64
63. Guo Z, Johnston W, Kovtun O, Mureev S, Brocker C, Ungermann C, Alexandrov K (2013) Subunit organisation of in vitro reconstituted HOPS and CORVET multisubunit membrane tethering complexes. PLoS One 8(12), e81534
64. Fernandez-Robledo JA, Vasta GR (2010) Production of recombinant proteins from protozoan parasites. Trends Parasitol 26(5):244–254
65. Fernandez FJ, Vega MC (2013) Technologies to keep an eye on: alternative hosts for protein production in structural biology. Curr Opin Struct Biol 23(3):365–373

Alternative Eukaryotic Expression Systems for the Production of Proteins and Protein Complexes

11

Sara Gómez, Miguel López-Estepa, Francisco J. Fernández, Teresa Suárez, and M. Cristina Vega

Abstract

Besides the most established expression hosts, several eukaryotic microorganisms and filamentous fungi have also been successfully used as platforms for the production of foreign proteins. Filamentous fungi and *Dictyostelium discoideum* are two prominent examples. Filamentous fungi, typically *Aspergillus* and *Trichoderma*, are usually employed for the industrial production of enzymes and secondary metabolites for food processing, pharmaceutical drugs production, and textile and paper applications, with multiple products already accepted for their commercialization. The low cost of culture medium components, high secretion capability directly to the extracellular medium, and the intrinsic ability to produce post-translational modifications similar to the mammalian type, have promoted this group as successful hosts for the expression of proteins, including examples from phylogenetically distant groups: humans proteins such as IL-2, IL-6 or epithelial growth factor; α-galactosidase from plants; or endoglucanase from *Cellulomonas fimi*, among others. *D. discoideum* is a social amoeba that can be used as an expression platform for a variety of proteins, which has been extensively illustrated for cytoskeletal proteins. New vectors for heterologous expression in *D. discoideum* have been recently developed that might increase the usefulness of this system and expand the range of protein classes that can be tackled. Continuous developments are ongoing to improve strains, promoters, production and downstream processes for filamentous fungi, *D. discoideum*, and other alternative eukaryotic hosts. Either for the overexpression of individual

S. Gómez • M. López-Estepa • F.J. Fernández
T. Suárez (✉) • M.C. Vega (✉)
Center for Biological Research, Spanish National Research Council (CIB-CSIC),
Ramiro de Maeztu 9, 28040 Madrid, Spain
e-mail: teresa@cib.csic.es; cvega@cib.csic.es

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_11

genes, or in the coexpression of multiples genes, this chapter illustrates the enormous possibilities offered by these groups of eukaryotic organisms.

Keywords
Filamentous fungi • *Aspergillus* • *Trichoderma* • *Hypocrea* • *Dictyostelium discoideum*

11.1 Introduction

The available repertoire of eukaryotic expression platforms has a tremendous potential for the production of highly complex proteins and protein assemblies, as has been amply demonstrated [1]. Besides the most established eukaryotic hosts that, like yeasts and insect cells, have been applied to a very wide range of protein sequences, there exist alternative hosts with unique properties that make them especially well tailored for the production of specific proteins. Two such eukaryotic hosts are dealt with in this chapter: filamentous fungi and the social amoeba *Dictyostelium discoideum*. They are phylogenetically diverse and have distinct biochemical, metabolic, cellular, and organismal properties; therefore they find applications in different fields. Most filamentous fungi, including species from the *Aspergillus* and *Trichoderma* genera, have been most often used in the context of industrial enzyme production, while *D. discoideum* has been harnessed for the production of cytoskeletal proteins.

11.2 Filamentous Fungi

Microorganisms classified as fungi (yeasts and molds) form a very diverse and complex group. Ascomycota, considered the largest phylum of the kingdom Fungi, comprises around 50 % of all currently known species and is subdivided into three subphyla: Saccharomycotina (including yeast species widely used such as *Saccharomyces cerevisiae* and *Candida albicans*), Taphrinomycotina (believed to be the most primitive of the three subphyla) and Pezizomycotina, which includes many species commonly referred to as filamentous fungi, such as *Aspergillus*, *Trichoderma* or *Penicillium*.

Numerous industrial products including proteins and enzymes are obtained using filamentous fungi as cell factories due to their inherently high protein secretory ability, glycosylation machinery to accomplish post-translational protein modifications, commercially available GRAS strains, and low cost of culture medium. Nevertheless, typical protein yields have remained low in comparison with mainstream hosts like *E. coli* or yeasts, which has motivated a surge of developments aimed at improving the overall performance of filamentous fungi as expression hosts (Fig. 11.1). Most efforts are directed towards the incorporation of genetically engineered traits that increase protein quality and overall yield.

Research on recombinant protein expression in filamentous fungi concentrates on two genera of ascomycetes, *Aspergillus* and *Trichoderma* (Table 11.1). Using them as model systems, many strategies have been implemented or are currently underway to enhance heterologous protein expression [2–7].

11.2.1 A Vastly Useful Genus: *Aspergillus*

Aspergillus is a diverse genus with high social and economic impact that comprises in excess of 335 mold species [14]. *Aspergillus* is most frequently observed as an anamorph stage (*i.e.*, it propagates by asexual reproduction), although some teleomorph forms have also been described. Starting in 1965, the classification of *Aspergillus* spp. was first pursued on the basis of morphological characters [15], but with the advent of

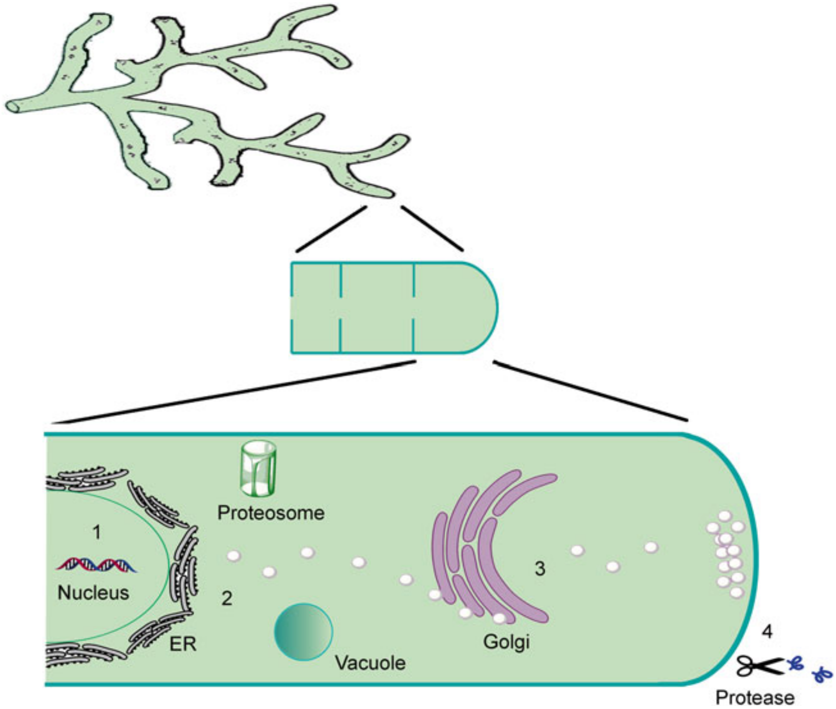


Fig. 11.1 Optimization of recombinant expression in filamentous fungi. Messenger RNA (mRNA) is synthesized in the nucleus and transported to the cytosol where translation into proteins occurs in the ribosomes associated with the endoplasmic reticulum (ER). Folding of proteins targeted for secretion and first glycosylations take place concurrently. Then, proteins in ER are transported for further post-translational modifications (*e.g.*, glycosylation) to the Golgi into secretory vesicles. Finally, modified proteins are carried by secretory vesicles to the hyphal tip for secretion. Improperly glycosylated or misfolded proteins are sent to the proteasome or vacuoles for degradation. Stages at which optimization for enhanced protein yields is possible in filamentous fungi are marked: (1) Vector design optimization. (2) Codon usage. (3) Glycosylation; quality control. (4) Protease inhibition

Table 11.1 Selected commercially available proteins and protein compounds which are produced industrially using filamentous fungi (only *Aspergillus* and *Trichoderma* examples are shown)

Compound	Host organism	Application area	Reference
Catalase	*A. niger*	Food industry	[8]
Cellulase	*A. oryzae*	Textile and paper industry	[9]
Cellulase	*T. reesei*	Textile and paper industry	[9]
β-galactosidase	*A. oryzae*	Food industry	[8]
α-glucanase	*T. reesei*	Food industry	[8]
Glucose oxidase	*A. niger*	Textile industry and biosensor	[10]
Phytase	*A. niger*, *A. oryzae*	Food industry	[8]
Xylanase	*A. niger*, *A. oryzae*, *T. reesei*	Textile, paper and bakery industry	[11]
Citric acid	*A. niger*	Food and beverage industry	[12]
Lovastatin	*A. terreus*	Pharmaceutical industry	[13]

DNA sequencing and molecular phylogeny from 1995 onwards many new species have been grouped in the *Aspergillus* genus [14]. Being highly aerobic and chemo-organotrophs, many *Aspergillus* species can thrive in a variety of oxygen-rich environments where they grow as molds, feeding on glucose and polysaccharide sources and colonizing many plants and trees. Nearly one quarter of *Aspergillus* species have been implicated in plant, animal and human diseases. Their rapid growth and simple nutrient requirements renders them susceptible to grow in shake flasks and bioreactors on cheap substrates. *Aspergillus* is an interesting model organism in biotechnology for its well-known ability to produce a wide range of industrially and pharmacology secondary metabolites (including antibiotics and organic acids) as well as secreted proteins, often with enzymatic activities, or because of its role in food fermentations. Some of the genetic and metabolic features that make *Aspergillus* spp. attractive microorganisms for foreign protein production derive from its lifestyle and eukaryotic nature: (i) They possess the machinery to catalyze mammalian-like posttranslational modifications (*e.g.*, glycosylation); and (ii) they have an inherently high-capacity secretory pathway. The most commonly used species are *A. niger*, *A. awamori*, *A. oryzae*, *A. nidulans* and *A. terreus*. In particular, *A. oryzae* and *A. niger* are on the Generally Recognized as Safe (GRAS) list of the Food and Drug Administration (FDA) in the United States. Some modifications may be further introduced to improve protein yields based on the rational alteration of the chief factors involved in gene expression, translation and secretion, such as vector design (*e.g.*, through marker and promoter selection among others elements), protein engineering by codon optimization, protease gene disruption (*e.g.*, using protease deficient strains), or by introducing glycosylation sites to enhance protein stability and secretion. These strategies can be complemented by careful adjustment of the composition of culture media and optimization of grown variables (*e.g.*, trying alternative carbon sources) [16, 17].

11.2.2 Trichoderma

Trichoderma is an ascomycete fungal genus comprising more than 150 different species [18] with a widespread distribution over the world's soils where they tend to be the most prevalent culturable fungus. Although *Trichoderma* species are fundamentally mycotrophic (mycoparasitic and saprotrophic) and have therefore been used as biocontrol fungi, some species can engage in asymptomatic (mutualistic) endophytic relationships and a few have been found to cause opportunistic infections in humans. The *Trichoderma* genus was first described in 1794, but it was not until 1865 that *Trichoderma* was shown to represent the sexual reproductive stages or teleomorphs of *Hypocrea*, which in turn have *Trichoderma* as their anamorphs [19, 20]. *Trichoderma* possesses some distinctive morphological traits, including a high growth rate, a characteristic repetitive conidiophore structure, and green conidia [21]. As in *Aspergillus*, rapid growth in inexpensive media, existence of eukaryotic post-translational modification machinery, and an inherently high secretory capacity make *Trichoderma* spp. interesting as expression hosts, especially of secreted enzymes. Some *Trichoderma* spp. that have been exploited as expression platforms including *T. atroviride, T. harzianum, T. virens*, *T. asperellum* and *T. reesei* (teleomorph: *Hypocrea jecorina*). Members of the Longibrachiatum clade of *Trichoderma* such as *T. longibrachiatum* and *T. reesei* are ubiquitous colonizers of cellulosic materials and stand as potential tools for biomass degradation because of their remarkable capacity to produce large amounts of enzymes with cellulose and hemicellulose hydrolytic activities. The gene expression program that controls cellulose hydrolysis in *Trichoderma* appears to be regulated in part by the effect of light, which would provide researchers with a convenient means of inducing the heterologous expression of target proteins [22]. In addition to its role in biocatalysis, *T. reesei* has aroused interest also for the bioconversion of waste organic matter into biofuels, and other *Trichoderma* species have been exploited for the production of secondary metabolites such as nonribosomal peptides, polyketides, isoprenoids and pyrones [23]. These applications explain why *T. reesei* has become the *Trichoderma* species of choice, with most strain and genome engineering research focused on its improvement. Indeed, several hyper-producing strains of *T. reesei* are commercially available such as QM9414 (catabolite repressed, cellulase hyperproducer strain from NatickLabs) and RUT-C30 (catabolite de-repressed, cellulase hyperproducer strain from Rutgers University) [24–26]. Both hyper-producer strains were obtained in the 70s from random mutagenesis of the wild type *T. reseei* QM6a with the aim of obtaining high levels of cellulolytic enzyme production for further applications in the search for fuel alternatives. Resulting cellulose activity showed an increase of 2–4 times for QM9414 and 15–20 times for RUT-C30, becoming this last strain as preferentially chosen. Despite its advantages, a few limitations of *Trichoderma* as host for foreign protein production remain, such as a differential glycosylation pattern and the production of many proteases that can degrade the expressed protein product. Attempts to express glycosylated mammalian proteins in *Trichoderma* had been limited owing to incomplete or extraneous *N*-glycosylation pattern. Although the core glycan is mammalian-like, glycosylated proteins are however expressed as non-sialylated and terminally decorated with non-mammalian sugars. Co-expression of the target mammalian glycoproteins with β-1,4-galactosyl transferase and α-2,6-sialyl transferase, two enzymes involved in mammalian *N*-glycosylation, has proven to be a viable strategy for the production of mammalianized glycosylated proteins in *Trichoderma* [7].

11.2.3 Genetic Engineering of Filamentous Fungi

Several genetic engineering strategies have been developed in recent years that have targeted the perceived bottlenecks and traditional limitations in the protein production pipeline using filamentous fungi. Research has focused on the optimization of transformation methodology, the search for promoters and selectable markers, the resolution of secretion and posttranslational problems, and the employment of mutant strains [5, 27–29].

11.2.3.1 Transformation Methods

Filamentous fungi like *Aspergillus* and *Trichoderma* are characterized by the presence of a thick cell wall and a low capacity to maintain plasmids. Development of efficient transformation methods is an indispensable prerequisite for successful strain engineering. A list of well-established transformation methods includes:

- Protoplast mediated method: relies on the uptake of DNA by fungal cells after enzymatic removal of the cell wall in the presence of high concentrations of polyethylene glycol (PEG). The principal disadvantage is the low efficiency with which protoplasts can be obtained. It is the most frequently used method [30].
- Electroporation: Cells can be reversibly made permeable to DNA by subjecting them to an electric pulse; optimum conditions require experimentation [31].
- *Agrobacterium tumefaciens*-mediated transformation: Developed initially for transformation of plants, transferring the Ti plasmid to fungi can be achieved in a similar way as in plants [32].
- Biolistic transformation: This technique depends on the high-speed bombardment of cells with DNA-coated metal particles [33].

Often, high frequency transformation in *Aspergillus* is associated with the genomic integration of the transformed plasmid DNA and, when multiple copies of an expression plasmid are integrated in the genome, the resultant expression levels can be greatly increased through a copy-number effect [34].

11.2.3.2 Vectors and Selectable Markers

Selectable markers are introduced into strain specific vectors to facilitate recognition of transformants. These markers can be divided into three groups:

- Dominant markers: they generally encode antibiotic resistance markers against chemical drugs (*hph* for Hygromycin B, *ble* for Phleomycin, *neo* for Neomycin, Geneticin and Kanamycin resistance, and *bar* for Glyphosate resistance) [35].
- Reporters: designed to facilitate visual differentiation (Gus, LacZ and eGFP) [36, 37].
- Complementation markers: introduction of functional genes, such as auxotrophic markers (*pyrG* for orotidine-5′-monophosphate decarboxylase, *hxk1* for hexokinase, *niaD* for nitrate reductase, and *argB* for ornithine carbamoyltransferase), nutritional marker genes (*CBS1* for cystathionine β-synthase, *HPD4* for p-hydroxyphenylpyruvate dioxygenase, *ptrA* for pyrithiamine resistance and, the most frequently used marker, *amdS* for acetamidase), and conditional lethal genes (*HSVtk* for herpes simplex virus thymidine kinase) [38, 39].

Combining several different marker types in a transformation can greatly increase the process of strain selection by preventing the unintended selection of false positive strains, *e.g.*, cotransforming a dominant resistance marker (antibiotic resistance cassette) together with a GFP marker, and selecting successfully transformed cells on selective medium under a confocal fluorescence microscope.

Vectors to provide high-level expression are constantly under development and typically incorporate in their multicloning sites sequences for oligopeptide tags for affinity purification (*e.g.*, oligohistidine tag, streptavidin affinity tag) that can be fused to the 5′ or 3′ ends of the gene of interest. Recombinant proteins obtained through this approach can be purified using affinity resins in a simple one-step process. Doubly tagging the gene of interest at either end with a different affinity tag has the additional advantage that the purified protein product will be free from proteolyzed protein fragments.

A choice exists between expression vectors that are nonintegrative (or episomal) and integrative plasmids. Episomal vectors are poorly maintained in filamentous fungi and tend to be used for the preliminary characterization of expression experiments. Conversely, integrative vectors direct the targeting of the expression cassette into a specific locus within the fungal genome, thus permitting the generation of very stable overproducing strains without the need to implement costly selection strategies. A drawback of integrative plasmids is the nonhomologous end joining (NHEJ) pathway, which may be very active in fungi in comparison with conventional homologous recombination and can mistarget expression cassettes into random genome locations where expression might be suboptimal. Reports of fungal strains where the Ku70-Ku80 loci have been knocked out show that most off-target integration events can be eliminated by impairing the normal function of the NHEJ cellular machinery [40–42].

11.2.3.3 Promoter Selection

Several strong constitutive and inducible promoters have been identified for use in filamentous fungi species. Tables 11.2 and 11.3 show a summary of the most common promoters used on *Aspergillus* and *Trichoderma* research. Commonly employed strategies for improvement of expressed protein yields are based on the selection of suitable promoters for each specific protein.

In *Aspergillus*, a wide range of heterologous proteins have been successfully expressed: human proteins (*e.g.*, IL-2, IL-6, epithelial growth factor), proteins from other animals (*e.g.*, porcine pancreatic phospholipase A2, hen egg-white lysozyme), plant proteins (*e.g.*, α-galactosidase from *Cyamosis tetragonoloba*), bacterial proteins (*e.g.*, endoglucanase from *Cellulomonas fimi*, *Clostridium thermocellum* dockerinc) and fungal proteins (*e.g.*, *Thermomyces lanuginosus* lipase, *Trametes versicolor* laccase) [17, 29, 43, 44].

In the case of *T. reesei*, most of the heterologous genes that have been overexpressed come from *Trichoderma* spp., with fewer attempts made to express heterologous genes using alternative fungi, such as cinnanoyl esterase from the

Table 11.2 Promoters used for protein expression in *Aspergillus*

Promoter	Gene regulated	Organism source	Constitutive/inducible
glaA	Glucoamylose	*A. niger*	Maltose, starch
alcA	Alcohol dehydrogenase	*A. nidulans*	Ethanol
alC	Alcohol dehydrogenase	*A. nidulans*	Ethanol
exlA	Endoxylanase	*A. awamori*	Xylose
thiA	Involved in thiamine biosynthesis	*A. oryzae*	Thiamine-dependent
aphA	Acid phosphatase	*A. nidulans*	Phosphate
sodM	Superoxide dismutase	*A. oryzae*	Addition of H_2O_2
gpdA	Glyceraldehyde-3-phosphate dehydrogenase	*A. nidulans*	Constitutive
adhA	Alcohol dehydrogenase	*A. nidulans*	Constitutive
tpiA	Triosephosphate isomerase	*A. nidulans*	Constitutive
pkiA	Protein kinase A	*A. oryzae*	Constitutive
gdhA	Glutamate dehydrogenase	*A. awamori*, *A. niger*	Constitutive
oliC	ATP synthase	*A. nidulans*	Constitutive
tef1	Translation elongation factor	*A. oryzae*	Constitutive
oliC/acuD	Hybrid promoter	Hybrid promoter	Acetate

Table 11.3 Promoter used for protein expression in *Trichoderma*

Promoter	Gene regulated	Organism source	Constitutive/inducible
cbh1	Cellobiohydrolase I	*T. reesei*	Cellulose, sophorose, lactose
cbh2	Cellobiohydrolase II	*T. reesei*	Cellulose, sophorose, lactose
xyn2	Xylanase	*T. reesei*	Cellulose, sophorose, lactose
egl2	Glycosyl hydrolase	*T. reesei*	Data not available
rp2	Ribosomal protein	*T. reesei*	Constitutive
pgk1	Pyruvate kinase	*T. reesei*	Constitutive
pkiA	Protein kinase A	*T. reesei*	Constitutive
pdC	Pyruvate decarboxylase	*T. reesei*	Constitutive
tef1	Translation elongation factor	*T. reesei*	Constitutive
eno	Enolase	*T. reesei*	Constitutive

unculturable anaerobic fungus *Piromyces equi* [4, 25, 44–48].

11.2.3.4 Codon Usage

In *Aspergillus*, as well as in other species of filamentous fungi, %G+C genome content is around 50 %. For that reason, codon optimization of the recombinant gene prior to transformation is not absolutely necessary for the overexpression of mammalian proteins, although it might still be advisable if the codon usage of the selected host significantly differs from that of the source organism. In *A. nidulans*, a group of 20 optimal codons have been established that are characterized by a G or C at the wobble position, with the explicit recommendation that an A at the third position should be avoided; in contrast, in humans CpG codons are underrepresented, most likely because they have been associated with mutational hotspots. Therefore, modifications in codon usage to overcome codon limitation are essential if human proteins were to be expressed [5, 49, 50]. In some cases, codon optimization by replacing rare codons by frequently used codons has had dramatic changes in the expression level. For example, Nelson et al. demonstrated this strategy for the expression of aequorin D (*aeqD* gene) from the jellyfish *Aequorea victoria* in *N. crassa*, *A. niger* and *A. awamori* in the context of the design of an intracellular Ca^{2+} sensor. Codon bias analysis suggested that as many as 44 out of the 197 codons in the wild-type *aeqD* gene were rarely used in the filamentous fungi. A synthetic version of the *aeqD* gene with corrected codon bias was placed under the control of the *Neurospora* clock-controlled gene (*ccg-1*) for overexpression in *N. crassa* under glucose starvation or under the control of the strong constitute glyceraldehyde phosphate dehydrogenase (*gpdA*) promoter of *A. nidulans* for overexpression in *A. niger* and *A. awamori*. The net increase in protein yields was 45 times more in *N. crassa* with respect to the wild-type *aeqD* gene expressed in the same promoter context, and 5 and 10 times greater than that in *A. niger* and *A. awamori* [51].

11.2.3.5 Secretion Pathway and Glycosylation

In filamentous fungi, translation and translocation into the endoplasmic reticulum (ER) occur concomitantly and additional modifications, such as further glycosylation and peptide processing, occur in Golgi bodies. Misfolded or improperly glycosylated proteins are sent to the proteasome or to vacuoles for degradation. Despite the capacity of filamentous fungi for secreting large amounts of proteins, improving the secretory pathway remains a major bottleneck for protein expression, therefore many efforts are dedicated to find ways to overcome this limitation [2, 5]. One successful strategy consists in fusing the gene of interest to the 3′ of a protein naturally secreted in high amounts (a “carrier”), which is thought to facilitate the transit of the foreign protein through the secretory pathway. A common carrier gene encodes for glucoamylase from *A. niger*, but other genes have also been used. An obvious disadvantage of this approach occurs when the carrier protein fusion needs to be proteolytically cleaved before the target protein can be used. To achieve this, the coding sequence for a recombinant protease must be introduced between the carrier and the target gene.

It is also worth mentioning that the secretion of overexpressed proteins in *Aspergillus* or *Trichoderma* can trigger the unfolding protein response (UPR), which blocks secretion and upregulates proteolytic activities in the cell. Despite much research, the molecular mechanisms behind the UPR remain poorly characterized and, therefore, inaccessible to genetic manipulation [7, 36, 52].

11.2.3.6 Secreted Proteases

Many heterologous proteins are proteolytically degraded during overexpression as a result of the unspecific action of proteases from the host organism; this problem is exacerbated in filamentous fungi due to the large amounts of proteases that they naturally secrete to the extracellular medium. A wide variety of proteases are encoded in the genome of filamentous fungi, with a broad range of optimal pH, thereby posing a significant challenge for successful protein production of protease-sensitive proteins. In *Aspergillus*, around 100–200 genes encode for a very diverse and species-specific group of proteases, which have been annotated based on available genome data. Two methods have been developed to overcome this limitation: optimization of culture medium composition, sometimes coupled with the choice of promoters that are most active under conditions that repress protease activity; and the engineering of protease-deficient strains as recombinant hosts, prepared through multiple rounds of random mutagenesis followed by selection or by targeted protease gene disruption [53, 54].

Protease activity is affected by the pH of culture medium. Strict control of culture conditions has been proposed as a useful strategy for decreasing secreted acid aspartic protease activity in *Aspergillus*. Buffering medium with a pH value near 6.0 resulted in limited acidic protease activity. However, at neutral pH only acidic proteases are inhibited, whereas alkaline and neutral proteases remain active [55]. Combining this approach with the selection of promoters that drive expression only under conditions that repress protease activity could enhance protein yield. The *pki*A promoter from protein kinase A, for example, allows constitutive expression at high glucose concentration and in presence of ammonium, precisely those culture conditions under which most fungal proteases are less active [56].

Isolating protease-deficient strains has been attempted in *A. niger* by randomly mutagenizing parental strain AB4.1 with UV irradiation, then applying selection methods that test extracellular protease activity. Thus, a mutant strain, designated AB1.13, was isolated with a residual protease activity of 1–2 % compared with AB4.1. Genomic analyses of mutations in AB1.13 have revealed that major protease expression, including *pep*A, which encodes the major acid protease in *Aspergillus*, were disrupted [57, 58]. Taking into consideration the complexity of protease regulation, disruption of the genes encoding proteases in filamentous fungi is no easy task. An approach that has revealed successful is the use of gene replacement techniques. Moralejo et al. have carried out an interesting work in *A. awamori*, with the aim to obtain a mutant in *pep*B, one of major extracellular proteases, through partial silencing by antisense mRNA. A plasmid for the recombinant expression of thaumatin, a sweet-tasting additive for food and feed, was constructed that incorporated the *pep*B gene in the antisense orientation under the control of the strong constitutive *gpd*A promoter. When this plasmid was transformed in *A. awamori*, a 31 % increase in expression of thaumatin in comparison with a control plasmid was observed. Since in this *A. awamori* recombinant strain residual protease activity could be detected, a double recombination with a plasmid bearing two selectable markers (hygromycin resistance and *pyr*G auxotrophy) was performed to knock out the *pep*B gene completely. Together with medium optimization, *pep*B knock-out strains exhibited an even greater increase in thaumatin expression approaching 90 % with respect to the control strain [59]. In addition to the *pep*A and *pep*B gene disruptions discussed above, disruption of the alkaline serine protease SPW in *Trichoderma* by insertional recombination together with pH controlled media have been applied by Zhang et al. in a recent work on a heterologous alkaline endoglucanase, showing a halving of the protease activity and a concomitant increase in the yield of alkaline endoglucanase [53].

11.2.4 Coexpression in Filamentous Fungi

11.2.4.1 Aspergillus

Although filamentous fungi are not the most common host in order to express recombinant proteins, their inherent capacity to secrete a large amount of protein (endogenous or recombinant) renders filamentous fungi as an interesting platform for the industry [60]. Some researchers have developed approaches to utilize *Aspergillus* and *Trichoderma* for the simultaneous coexpression of multiple proteins [7].

An imposing limitation commonly observed in filamentous fungi for the secretion of overexpressed proteins is the trigger of the UPR stress response [61, 62]. Under certain circumstances, coexpression of recombinant proteins with chaperones can increase the amount of secreted protein [63]. For instance, Conesa et al. [62] were able to coexpress manganese peroxidase (MnP1) from *Phanerochaete chrysosporium* (an enzyme of interest because of its ability to degrade lignin), with the help of two chaperones independently, calnexin (CLX) and binding protein (BiP), using *A. niger* strain MGG026 as host [prtT gla::fleo^r pyrG] [64] (Table 11.4). This strain was first transformed with the vector pgpdMnP1.I-AmdS, constructed with the *MnP1* gene under the control of the strong promoter of the glyceraldehyde-3-phosphate dehydrogenase from *A. nidulans*, and the *amdS* gene from *A. nidulans* as selection marker, which allows *A. niger* to metabolize acetamide as a carbon and nitrogen source [65]. The calnexin gene was cloned under the *A. niger* glucoamylase promoter to generate pGLACLX, which was transformed in combination with pAN7-1. For the *bip* gene, the same vector was used, but with the addition of hygromycin resistance as selection marker.

Firstly, *A. niger* was cotransformed with pgpdMnP1.I-AmdS and pAB4-1 [66], followed by the screening of peroxidase activity at 30 °C to select strains with acceptable expression levels of MnP1 [64]. Secondly, a previously selected strain was cotransformed with pGLABiP/hph or

Table 11.4 Summary of plasmids used in Conesa et al. [62] to increase MnP1 secretion by coexpression of calnexin or BiP. *gpdA* glyceraldehyde-3-phosphate dehydrogenase promoter, *mnpI* manganese peroxidase, *Amp* ampicillin resistance, *amdS* acetamidase, *pyrG* orotidine-5′-phosphate decarboxylase, *hph* hygromycin resistance, *bip* binding protein, *clxA* calnexin, glaA glucoamylase promoter

Vector	Constitutive elements	Markers	Step
pgpdMnP1-I-AmdS	*gpdA* promoter, *mnpI* gene	Amp, amdS	First transformation of MGG029
pAB4-1	–	*pyrG*	
pGLABiP/hph	*glaA* promoter, *bip* gene	Amp, hph	Cotransformation of selected MGG029 from step I
pGLACLX	*glaA* promoter, *clxA* gene	Amp	Cotransformation of selected MGG029 from step I
pAN7-1	–	hph	

pGLACLX plus pAN7-1, both systems allowing selection by hygromycin [62, 67]. The assay for peroxidase activity showed significantly higher yield in the case of coexpression with the construct that carries the calnexin. Selected strains cotransformed with both chaperones were employed in a following experiment to test the overproduction in shake flask cultures using AMM medium (*Aspergillus* maltodextrin medium, in order to start the induction) or AMM supplemented with hemin, to facilitate the secretion of MnP as a control. Results showed, again, that the yield was higher with coexpression with calnexin, thereby concluding that this chaperone helps the correct folding and secretion of the protein of interest [62]. Despite this success, the authors suggest that not every recombinant protein secreted by *A. niger* might be improved only by coexpressing with one chaperone, as they showed from the negative result with BiP.

Coexpression of multiple proteins in filamentous fungi has also been used for the production of secondary metabolites and small chemical compounds and, additionally, to produce enzymatically modified versions of those molecules for specific purposes. A significant example lies in the generation of recombinant *A. nidulans* strains that overproduce brevianamide F [68], a prenylated non-ribosomal peptide (NRP) not present naturally in *Aspergillus* (Fig. 11.2). The incorporation of novel synthetic pathways comprising non-ribosomal peptide synthetases (NRPS) and various prenyltransferases was required for the assembly of a working brevianamide F biosynthetic pathway. In this case, the strain employed for the coexpression was *A. nidulans* TN02A7 (pyrG89, pyroA4, nkuA::argB; riboB2) [42]. The gene *tfmPS* (from *Neosartorya fischeri)* was cloned in pJW24 (which possesses the *pyrG* gene), which was under the control of the strong glyceraldehyde-3-phosphate dehydrogenase (*gdpA*) promoter in the vector pCAW28. The *tfmPS* gene was coexpressed with three different prenyltransferases, two of them from *N. fischeri* (*cdpC2PT* and *cdpC3PT*) and one from *A. fumigatus* (*cdpNPT*). The gene *cdpNPT* was cloned in the vector pCaW34stop which contains sequences for the selection marker *pyroA* (used in media lacking pyridoxine), the *gdpA* promoter and the *trpC* terminator. The plasmids that code for *cdpC2PT* and *cdpC3PT* have the same characteristics as pCaW34stop, and were called pKM37 and pMK39 respectively. All these prenyltransferases acted over the molecule cyclo-L-Trp-L-Pro although in different ways to produce the end metabolite. Transformed strains were plated on media supplemented with uracil, uridine, riboflavin and pyridoxine, and grown in AMM to express the synthetic pathway and accumulate the prenylated non-ribosomal peptides. Analysis of the produced peptides was carried out by HPLC, nuclear magnetic resonance (NMR) and electron ionization mass spectrum (EIMS).

11.2.4.2 Trichoderma

Being used for the production of cellulases and hemicellulases [24] that are relevant for the development of biofuels, *Trichoderma reesei* has benefited from co-expression strategies intended

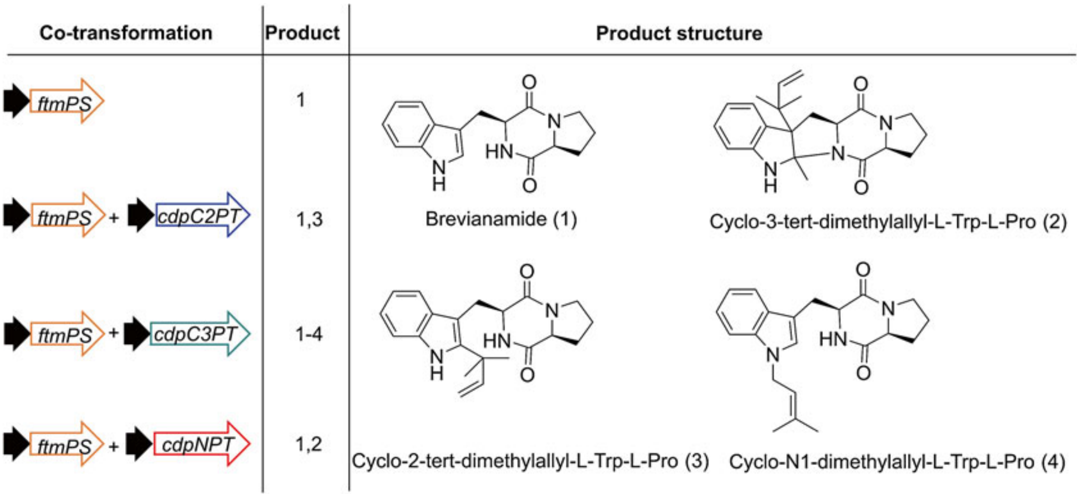

Co-transformation
Product
Product structure
ftmPS
1
ftmPS + cdpC2PT
1,3
ftmPS + cdpC3PT
1-4
ftmPS + cdpNPT
1,2
Brevianamide (1)
Cyclo-3-tert-dimethylallyl-L-Trp-L-Pro (2)
Cyclo-2-tert-dimethylallyl-L-Trp-L-Pro (3)
Cyclo-N1-dimethylallyl-L-Trp-L-Pro (4)

Fig. 11.2 The brevianamide F biosynthetic pathway: heterologous co-expression in *Aspergillus*. Schematic representation of a heterologous biosynthetic pathway designed to obtain a variety of prenylated non-ribosomal peptides. The relevant pathway enzymes are expressed in *A. nidulans* after co-transformation with a non-ribosomal peptide synthetase from *Neosartorya fischeri* (*ftmPS*) and three different prenyltransferases from *N. fischeri* and *A. fumigatus* [reverse C2-prenyltransferase gene (cdpC2PT), reverse C3-prenyltransferase gene (cdpNPT) and reverse C3-prenyltransferase gene (cdpC3PT)] [68]

to improve the production levels and secretion efficiency of recombinant *T. reesei* strains. One strategy exploits the dolichol-phosphate-mannose synthase gene from *Saccharomyces cerevisiae (DPM1)*, which encodes mannosylphosphodolichol (MPD) synthase [69, 70]. This gene is well known to play an important role in *O*-glycosylation in *T. reesei*, and its overexpression can achieve a significant improvement in protein secretion [71]. Also, choline was known to increase the yield of extracellular protein when added to *T. reesei* cultures, likely through the stimulation of MPD synthase activity [72]. The *DPM1* gene was cloned under the control of the pyruvate kinase (*pki1*) promoter and transformed into *T. reesei* TU-6 strain, which is auxotrophic for uridine, along with a plasmid bearing the *pyr4* gene to complement uridine auxotrophy and facilitate selection [73, 74]. Mitotic stability of the recombinant strains was analyzed to ensure the appropriate maintenance of the cotransformed plasmids [75, 76]. To quantitate the expression of MPD synthase, a radioactive assay of *T. reesei* membrane fraction was established that monitored the transfer of mannose residues across the membrane [77]. This carefully constructed recombinant strain was then tested as to its potential to sustain high-level secretion of cellobiohydrolase I (CBHI), an exocellulase that accounts for roughly 50 % of the total *T. reesei* secretome, clearly showing by comparison with a negative control strain (not transformed with the *DPM1* plasmid) that co-expression indeed increases CBHI secretion. As expected, this enhanced secretion of CBHI was independent of the amount of *cbh1* mRNA, which was measured by Northern blotting, thereby suggesting that the improved secretion of CBHI must be due to a post-translational effect exerted or mediated by the overexpressed MPD synthase [69].

Enzymes tailored for biofuel production are typically required to have high activity over a broad pH range. Wang et al. [78] made an exhaustive search for the optimal vector construct to express simultaneously various proteins in order to achieve cellulose activity at basic pH, overcoming the limitation that *T. reesei* cellulases are more active at acidic pH [79]. Authors chose the RUT-C30 strain (ATCC 56765), a mutant able to hypersecrete cellulase [26], and used plasmids based on the pPK2 plasmid, with a hygromycin resistance gene placed upstream of the gene or

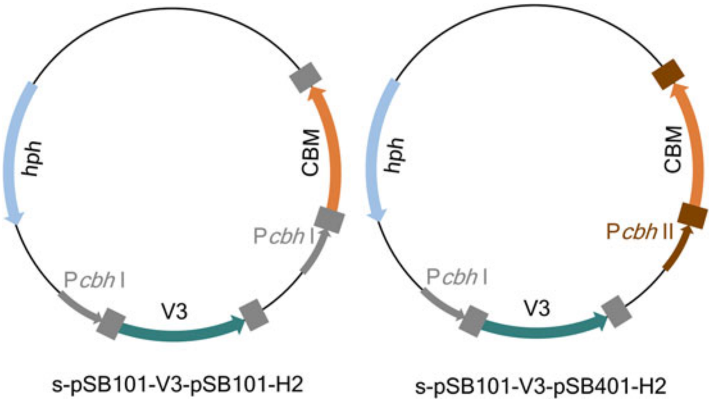


Fig. 11.3 In search for the perfect cellulase cocktail: heterologous co-expression in *Trichoderma*. Plasmids constructed for coexpression of the chimeric protein (encoded in the *egv3* gene) and the alkaline cellobiohydrolase (CBH2) in *T. reesei. Hph* hygromycin resistance, *egv3* gene encoding for chimeric protein, *CBH2* alkaline cellobiohydrolase from *H. insolens. Grey box and arrow* represent *cbh1* promoter, signal peptide-encoding sequence, and terminator; *brown box* and *arrow* represent *cbh2* promoter, signal peptide and terminator [78]

genes of interest. Of the enzymes produced using co-expression with this approach is the core enzyme of EGV from *Humicola insoles* (an endoglucanase whose maximum activity is at pH 7–9) fused to the carbohydrate-binding module (CBM) from *Humicola grisea var. thermoidea* (Fig. 11.3). This chimeric construct was finally selected as optimal cellulase to be coexpressed with an alkaline cellobiohydrolase (exoglucanase from *Humicola insolens*). Each gene is flanked by its own promoter and signal peptide-encoding DNA sequence at the 5′ to the gene and a terminator at the 3′ [78, 80]. Two constructs were developed: s-pSB101-V3-pSB101-H2, bearing the genes encoding the chimeric protein and the alkaline cellulase from *H. insoles*, each flanked with its own *cbh1*-derived promoter, signal sequence and terminator; and s-pSB101-V3-pSB401-H2, where the gene for the chimeric protein was flanked by *cbh1*-derived promoter, signal sequence and terminator, and the alkaline exoglucanase gene, flanked by *cbh2*-derived sequences [78]. These plasmids were transformed in UT-C30 by the ATM (*Agrobacterium tumefaciens*-mediated transformation) method [81] and cellulase activity secreted into the medium by the recombinant strains was monitored with two enzymatic assays: CMCase (Carboxymethyl Cellulose) activity and FPAase activity (Filter Paper Activity), at pH 5–9 at 50 °C. In both assays, the release of glucose from the cellulose substrate is monitored with the help of DNS (dinitrosalicylic acid) [82]. The outcome of these two assays showed that the s-pSB101-V3-pSB401-H2 construct resulted in the greatest activity at pH 8.0, whereas control strains and s-pSB101-V3-pSB101-H2 showed much reduced cellulase activity. This proved that the co-expression of the chimeric protein and the alkaline cellobiohydrolase created an active cellulase mixture capable of performing well at the neutral-basic pH ranges necessary for industrial applications.

11.3 *Dictyostelium discoideum*, Using a Social Amoeba for Heterologous Protein Expression

Dictyostelium discoideum is a eukaryotic organism that naturally lives as an amoeba in the soil of temperate forest, feeds by phagocytosis of microorganisms, like bacteria or yeast, and duplicates by cell division [83]. The structure and organization of the cell are more reminiscent of larger eukaryotic cells than other eukaryotic microorganisms like yeast, with a lipid bilayer plasma membrane, similar nuclear organization and cell cytoskeleton to metazoan cells [84]. Genome sequencing demonstrated that *D. discoideum* branched off from the main eukaryotic trunk after fungi, and before plants and animals separated in evolution. *D. discoideum* compact chromosomes have a high density of genes that encode around 12,500 predicted proteins [85]. This large number of genes, twice the number present in the yeast *Saccharomyces cerevisiae*, also a free living microorganism, and close to the number of genes present in the fly [85], is needed for this simple unicellular eukaryote to perform a sophisticated multicellular cycle that allows *D. discoideum* survival under difficult environmental conditions [83]. When there is no food available, *Dictyostelium* cells start to emit cAMP pulses to establish a cAMP relay that mediates intercellular communication and starts aggregation (Fig. 11.4A). At this stage, about 100,000 cells gather together to form a mound and initiate a tightly regulated developmental cycle (Fig. 11.4B). Cell differentiation to prestalk and prespore cell types has started during aggregation and cells now crowd together to form a mound. The mound will elongate and form, first a finger and later, a slug, which is motile and can migrate. In the final part of the cycle, the slug will form a fruiting body, with a ball of mature spores at the top of a stalk, made up of dead cells (Fig. 11.4B) [86].

Dictyostelium discoideum is a non-mammalian model organism for functional analysis of genes and proteins, approved by the National Institutes of Health (Bethesda, MD, USA). Many cellular and molecular aspects of its life cycle have been studied in detail and carefully dissected. Basic principles for cell-to-cell communication, cytoskeletal organization, gradient sensing and intracellular signaling have derived from studies with *Dictyostelium* [87], because the molecular machinery that controls fundamental aspects of chemotaxis, gradient sensing and phagocytosis are evolutionarily conserved between human blood cells and *Dictyostelium* [88]. *D. discoideum* is also emerging as a very powerful model system for investigating phagocytosis and the mechanisms of bacterial virulence, to explore the molecular basis of human diseases, as well as the mechanisms of drug action and the biochemical pathways that lead to resistance to certain therapeutic agents in human cells [89]. Most of these investigations have been fostered by the fact that the *D. discoideum* genome shows a higher degree of gene conservation with the human genome than with the fungal ones [85].

D. discoideum amoebae can be manipulated with much the same simplicity and low cost as bacteria or yeast in the laboratory; when they grow on axenic culture media, a large number of cells can be easily obtained in a few days [90].

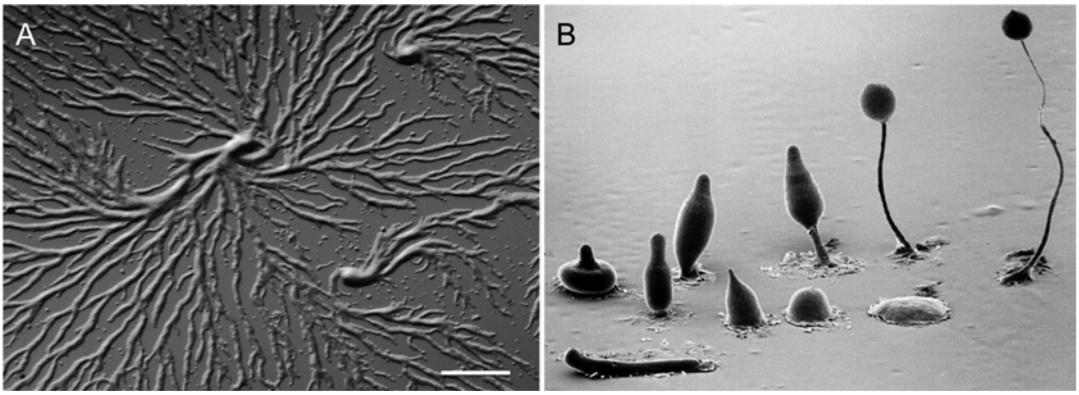


Fig. 11.4 *Dictyostelium discoideum* life cycle. **A** *D. discoideum* cell streaming after 8 h of aggregation, bar 1 mm (A. Garciandia and T. Suárez). **B** SEM of *D. discoideum* developmental structures (Reproduced with permission from M.J. Grimson and R.L. Blanton)

Foreign DNA can be efficiently delivered to *D. discoideum* cells by electroporation and selection of transformants can be achieved in 5–10 days. The availability of a wide variety of versatile vectors adapted for use in *Dictyostelium*, together with a reasonable number of different promoters (constitutive, regulatable, inducible), resistance markers and protein tags allows efficient protein expression in the amoeba [90, 91].

D. discoideum is an attractive system for heterologous expression for several classes of proteins and can properly fold complex heterologous proteins that are glycosylated. Its ability to perform post-translational modifications together with the high secretion level that *D. discoideum* can achieve, has allowed the production of glycoproteins [92], secreted growth factors [93] and complex recombinant proteins of therapeutic relevance [90] in large quantities and with low cost. Expression of human receptors that should be inserted in the plasma membrane to perform their correct function has also been achieved in this amoeba [94]. The possibility of easily generate randomly mutagenized proteins and screen for a particular characteristic or trait, provides a powerful tool for the analysis of protein function and the selection of new activities [95].

Dictyostelium genome has a very high A+T content, over 75 %, with non-coding DNA tracks that can reach 99 % [85]. This fact inflicts a clear bias in the codon preference and the presence of clusters of infrequent codons in the heterologous gene, particularly at the 5′ end of the coding sequence, can reduce or even abolish protein expression. Optimization of the first codons (10–15) of the heterologous gene before attempting expression in *D. discoideum* will generally improve protein expression levels [96]. A second parameter that can also improve protein expression is the assembly of the ribosome to the start codon, thus to conform the RNA sequence of the initiation site of translation to the *D. discoideum* consensus will definitively improve production [96].

To understand the mechanisms underlying some cellular and molecular processes that are preserved through evolution and present in *D. discoideum*, like cell chemotaxis or phagocytosis, the analysis of the structure of the proteins involved and the interactions among them is essential. Protein expression in the model system *D. discoideum* has started to contribute to the understanding of cytoskeleton flexibility and architecture [97, 98] and will undoubtedly be an extremely useful tool in the future with the outstanding bioinformatics support and common resources available to the *Dictyostelium* researchers [99, 100].

11.4 Conclusions

Filamentous fungi and *D. discoideum* have unique genetic and metabolic properties that make them unique in their capacity to produce large amounts of functional proteins and protein complexes that would otherwise be difficult to produce with other established methods. Filamentous fungi (especially *Aspergillus* and *Trichoderma*) have very efficient secretory machinery that makes them ideally suited for the production of extracellular enzymes, many of which possess biotechnological or therapeutic properties. *D. discoideum* occupies the opposite end of the expression host spectrum. This social amoeba's genome bears a closer resemblance to mammalian genomes than fungal genomes do, and as a consequence many cellular processes in higher eukaryotes can be modeled in *D. discoideum*. This is particularly true of cell motility and cytoskeletal protein complexes, since *Dictyostelium* has evolved a complex social organization that relies in the concerted movement and aggregation of a multitude of cellular individuals. Recombinant expression of cytoskeletal protein complexes has been successfully accomplished in *D. discoideum*, paving the way for further developments on other protein classes. Together, filamentous fungi and *D. discoideum* offer unique opportunities owing to peculiarities in their genomes and life styles that, wisely exploited, can succeed where more conventional hosts fail. They also represent examples of eukaryotic organisms that can be harnessed for targeted protein production.

Acknowledgments We gratefully acknowledge the financial support by Ministerio de Economía y Competitividad (ES) (grants PET2008_0101, BIO2009-11184, BFU2010-22260-C02-02 and CTQ2015-66206-C2-2-R to MCV) and the European Commission (Framework Programme 7) project ComplexINC No. 279039 to MCV.

References

1. Fernandez FJ, Vega MC (2013) Technologies to keep an eye on: alternative hosts for protein production in structural biology. Curr Opin Struct Biol 23(3):365–373
2. Conesa A, Punt PJ, van Luijk N, van den Hondel CA (2001) The secretion pathway in filamentous fungi: a biotechnological view. Fungal Genet Biol 33(3):155–171
3. Radzio R, Kück U (1997) Synthesis of biotechnologically relevant heterologous proteins in filamentous fungi. Proccess Biochem 32(6):529–539
4. Meyer V (2008) Genetic engineering of filamentous fungi – progress, obstacles and future trends. Biotechnol Adv 26(2):177–185
5. Su X, Schmitz G, Zhang M, Mackie RI, Cann IK (2012) Heterologous gene expression in filamentous fungi. Adv Appl Microbiol 81:1–61
6. Archer DB (2000) Filamentous fungi as microbial cell factories for food use. Curr Opin Biotechnol 11(5):478–483
7. Nevalainen KM, Te'o VS, Bergquist PL (2005) Heterologous protein expression in filamentous fungi. Trends Biotechnol 23(9):468–474
8. Olempska-Beer ZS, Merker RI, Ditto MD, DiNovi MJ (2006) Food-processing enzymes from recombinant microorganisms – a review. Regul Toxicol Pharmacol: RTP 45(2):144–158
9. Adrio JL, Demain AL (2003) Fungal biotechnology. Int Microbiol: Off J Span Soc Microbiol 6(3):191–199
10. Mirón J, Vázquez JA, González P, Murado MA (2010) Enhancement glucose oxidase production by solid-state fermentation of Aspergillus niger on polyurethane foams using mussel processing wastewaters. Enzym Microb Technol 46(1):21–27
11. Polizeli ML, Rizzatti AC, Monti R, Terenzi HF, Jorge JA, Amorim DS (2005) Xylanases from fungi: properties and industrial applications. Appl Microbiol Biotechnol 67(5):577–591
12. Dhillon GS, Brar SK, Verma M, Tyagi RD (2011) Utilization of different agro-industrial wastes for sustainable bioproduction of citric acid by Aspergillus niger. Biochem Eng J 54(2):83–92
13. Kumar S, Srivastava N, Gupta BS, Kuhad RC, Gomes J (2014) Lovastatin production by Aspergillus terreus using lignocellulose biomass in large scale packed bed reactor. Food Bioprod Process 92(4):416–424
14. Samson RA, Visagie CM, Houbraken J, Hong SB, Hubka V, Klaassen CH, Perrone G, Seifert KA, Susca A, Tanney JB, Varga J, Kocsube S, Szigeti G, Yaguchi T, Frisvad JC (2014) Phylogeny, identification and nomenclature of the genus Aspergillus. Stud Mycol 78:141–173
15. Raper KBF, Fennell DI (1965) The genus Aspergillus. Williams & Wilkins, Baltimore
16. Liu L, Liu Y, Shin HD, Chen RR, Wang NS, Li J, Du G, Chen J (2013) Developing Bacillus spp. as a cell factory for production of microbial enzymes and industrially important biochemicals in the context of systems and synthetic biology. Appl Microbiol Biotechnol 97(14):6113–6127
17. Lubertozzi D, Keasling JD (2009) Developing Aspergillus as a host for heterologous expression. Biotechnol Adv 27(1):53–75
18. Samuels GJ, Ismaiel A, Mulaw TB, Szakacs G, Druzhinina IS, Kubicek CP, Jaklitsch WM (2012) The longibrachiatum clade of trichoderma: a revision with new species. Fungal Divers 55(1):77–108
19. Person CH (1794) Disposita mehodica fungorum. Römer's N Mag Bot 1:81–128
20. Tulasne LR, Tulasne C (1865) Selecta fungorum carpologia. Jussu, Paris
21. Samuels GJ (1996) Trichoderma: a review of biology and systematics of the genus. Mycol Res 100(8):923–935
22. Schmoll M, Franchi L, Kubicek CP (2005) Envoy, a PAS/LOV domain protein of Hypocrea jecorina (Anamorph Trichoderma reesei), modulates cellulase gene transcription in response to light. Eukaryot Cell 4(12):1998–2007
23. Druzhinina IS, Seidl-Seiboth V, Herrera-Estrella A, Horwitz BA, Kenerley CM, Monte E, Mukherjee PK, Zeilinger S, Grigoriev IV, Kubicek CP (2011) Trichoderma: the genomics of opportunistic success. Nat Rev Microbiol 9(10):749–759
24. Schuster A, Schmoll M (2010) Biology and biotechnology of Trichoderma. Appl Microbiol Biotechnol 87(3):787–799
25. Singh A, Taylor LE 2nd, Vander Wall TA, Linger J, Himmel ME, Podkaminer K, Adney WS, Decker SR (2015) Heterologous protein expression in Hypocrea jecorina: a historical perspective and new developments. Biotechnol Adv 33(1):142–154
26. Peterson R, Nevalainen H (2012) Trichoderma reesei RUT-C30 – thirty years of strain improvement. Microbiology 158(Pt 1):58–68
27. Unkles SE, Valiante V, Mattern DJ, Brakhage AA (2014) Synthetic biology tools for bioprospecting of natural products in eukaryotes. Chem Biol 21(4):502–508
28. Mach RL, Zeilinger S (2003) Regulation of gene expression in industrial fungi: Trichoderma. Appl Microbiol Biotechnol 60(5):515–522
29. Pahirulzaman KA, Williams K, Lazarus CM (2012) A toolkit for heterologous expression of metabolic

pathways in Aspergillus oryzae. Methods Enzymol 517:241–260

30. Dawe AL, Willins DA, Morris NR (2000) Increased transformation efficiency of Aspergillus nidulans protoplasts in the presence of dithiothreitol. Anal Biochem 283(1):111–112
31. Ozeki K, Kyoya F, Hizume K, Kanda A, Hamachi M, Nunokawa Y (1994) Transformation of intact Aspergillus niger by electroporation. Biosci Biotechnol Biochem 58(12):2224–2227
32. Gouka RJ, Gerk C, Hooykaas PJ, Bundock P, Musters W, Verrips CT, de Groot MJ (1999) Transformation of Aspergillus awamori by Agrobacterium tumefaciens-mediated homologous recombination. Nat Biotechnol 17(6):598–601
33. Herzog RW, Daniell H, Singh NK, Lemke PA (1996) A comparative study on the transformation of Aspergillus nidulans by microprojectile bombardment of conidia and a more conventional procedure using protoplasts treated with polyethyleneglycol. Appl Microbiol Biotechnol 45(3):333–337
34. Miyauchi S, Te'o VS Jr, Bergquist PL, Nevalainen KM (2013) Expression of a bacterial xylanase in Trichoderma reesei under the egl2 and cbh2 glycosyl hydrolase gene promoters. New Biotechnol 30(5):523–530
35. Weld RJ, Plummer KM, Carpenter MA, Ridgway HJ (2006) Approaches to functional genomics in filamentous fungi. Cell Res 16(1):31–44
36. Ward OP (2012) Production of recombinant proteins by filamentous fungi. Biotechnol Adv 30(5):1119–1139
37. Jorgensen MS, Skovlund DA, Johannesen PF, Mortensen UH (2014) A novel platform for heterologous gene expression in Trichoderma reesei (Teleomorph Hypocrea jecorina). Microb Cell Factories 13(1):33
38. Jin FJ, Maruyama J, Juvvadi PR, Arioka M, Kitamoto K (2004) Development of a novel quadruple auxotrophic host transformation system by argB gene disruption using adeA gene and exploiting adenine auxotrophy in Aspergillus oryzae. FEMS Microbiol Lett 239(1):79–85
39. Navarrete K, Roa A, Vaca I, Espinosa Y, Navarro C, Chavez R (2009) Molecular characterization of the niaD and pyrG genes from Penicillium camemberti, and their use as transformation markers. Cell Mol Biol Lett 14(4):692–702
40. Jiang D, Zhu W, Wang Y, Sun C, Zhang KQ, Yang J (2013) Molecular tools for functional genomics in filamentous fungi: recent advances and new strategies. Biotechnol Adv 31(8):1562–1574
41. Takahashi T, Masuda T, Koyama Y (2006) Enhanced gene targeting frequency in ku70 and ku80 disruption mutants of Aspergillus sojae and Aspergillus oryzae. Mol Genet Genom: MGG 275(5):460–470
42. Nayak T, Szewczyk E, Oakley CE, Osmani A, Ukil L, Murray SL, Hynes MJ, Osmani SA, Oakley BR (2006) A versatile and efficient gene-targeting system for Aspergillus nidulans. Genetics 172(3):1557–1566
43. Punt PJ, van Biezen N, Conesa A, Albers A, Mangnus J, van den Hondel C (2002) Filamentous fungi as cell factories for heterologous protein production. Trends Biotechnol 20(5):200–206
44. Wang L, Ridgway D, Gu T, Moo-Young M (2005) Bioprocessing strategies to improve heterologous protein production in filamentous fungal fermentations. Biotechnol Adv 23(2):115–129
45. He R, Zhang C, Guo W, Wang L, Zhang D, Chen S (2013) Construction of a plasmid for heterologous protein expression with a constitutive promoter in Trichoderma reesei. Plasmid 70(3):425–429
46. Li J, Wang J, Wang S, Xing M, Yu S, Liu G (2012) Achieving efficient protein expression in Trichoderma reesei by using strong constitutive promoters. Microb Cell Factories 11:84
47. Penttila M, Nevalainen H, Ratto M, Salminen E, Knowles J (1987) A versatile transformation system for the cellulolytic filamentous fungus Trichoderma reesei. Gene 61(2):155–164
48. Zou G, Shi S, Jiang Y, van den Brink J, de Vries RP, Chen L, Zhang J, Ma L, Wang C, Zhou Z (2012) Construction of a cellulase hyper-expression system in Trichoderma reesei by promoter and enzyme engineering. Microb Cell Factories 11:21
49. Lloyd AT, Sharp PM (1991) Codon usage in Aspergillus nidulans. Mol Gen Genet MGG 230(1-2):288–294
50. Nabiyouni M, Prakash A, Fedorov A (2013) Vertebrate codon bias indicates a highly GC-rich ancestral genome. Gene 519(1):113–119
51. Nelson G, Kozlova-Zwinderman O, Collis AJ, Knight MR, Fincham JR, Stanger CP, Renwick A, Hessing JG, Punt PJ, van den Hondel CA, Read ND (2004) Calcium measurement in living filamentous fungi expressing codon-optimized aequorin. Mol Microbiol 52(5):1437–1450
52. Fleissner A, Dersch P (2010) Expression and export: recombinant protein production systems for Aspergillus. Appl Microbiol Biotechnol 87(4):1255–1270
53. Zhang G, Zhu Y, Wei D, Wang W (2014) Enhanced production of heterologous proteins by the filamentous fungus Trichoderma reesei via disruption of the alkaline serine protease SPW combined with a pH control strategy. Plasmid 71:16–22
54. Wiebe MG (2003) Stable production of recombinant proteins in filamentous fungi – problems and improvements. Mycologist 17(3):140–144
55. Eneyskaya EV, Kulminskaya AA, Savel'ev AN, Savel'eva NV, Shabalin KA, Neustroev KN (1999) Acid protease from Trichoderma reesei: limited proteolysis of fungal carbohydrases. Appl Microbiol Biotechnol 52(2):226–231
56. Wiebe MG, Karandikar A, Robson GD, Trinci AP, Candia JL, Trappe S, Wallis G, Rinas U, Derkx PM, Madrid SM, Sisniega H, Faus I, Montijn R, van den

Hondel CA, Punt PJ (2001) Production of tissue plasminogen activator (t-PA) in Aspergillus niger. Biotechnol Bioeng 76(2):164–174

57. Punt PJ, Schuren FH, Lehmbeck J, Christensen T, Hjort C, van den Hondel CA (2008) Characterization of the Aspergillus niger prtT, a unique regulator of extracellular protease encoding genes. Fungal Genet Biol 45(12):1591–1599
58. Mattern IE, van Noort JM, van den Berg P, Archer DB, Roberts IN, van den Hondel CA (1992) Isolation and characterization of mutants of Aspergillus niger deficient in extracellular proteases. Mol Gen Genet 234(2):332–336
59. Moralejo FJ, Cardoza RE, Gutierrez S, Lombrana M, Fierro F, Martin JF (2002) Silencing of the aspergillopepsin B (pepB) gene of Aspergillus awamori by antisense RNA expression or protease removal by gene disruption results in a large increase in thaumatin production. Appl Environ Microbiol 68(7):3550–3559
60. Archer DB, Peberdy JF (1997) The molecular biology of secreted enzyme production by fungi. Crit Rev Biotechnol 17(4):273–306
61. Gouka RJ, Punt PJ, van den Hondel CA (1997) Efficient production of secreted proteins by Aspergillus: progress, limitations and prospects. Appl Microbiol Biotechnol 47(1):1–11
62. Conesa A, Jeenes D, Archer DB, van den Hondel CA, Punt PJ (2002) Calnexin overexpression increases manganese peroxidase production in Aspergillus niger. Appl Environ Microbiol 68(2):846–851
63. Lombrana M, Moralejo FJ, Pinto R, Martin JF (2004) Modulation of Aspergillus awamori thaumatin secretion by modification of bipA gene expression. Appl Environ Microbiol 70(9):5145–5152
64. Conesa A, van den Hondel CA, Punt PJ (2000) Studies on the production of fungal peroxidases in Aspergillus niger. Appl Environ Microbiol 66(7):3016–3023
65. Kelly JM, Hynes MJ (1985) Transformation of Aspergillus niger by the amdS gene of Aspergillus nidulans. EMBO J 4(2):475–479
66. van Hartingsveldt W, Mattern IE, van Zeijl CM, Pouwels PH, van den Hondel CA (1987) Development of a homologous transformation system for Aspergillus niger based on the pyrG gene. Mol Gen Genet 206(1):71–75
67. Punt PJ, van den Hondel CA (1992) Transformation of filamentous fungi based on hygromycin B and phleomycin resistance markers. Methods Enzymol 216:447–457
68. Wunsch C, Mundt K, Li SM (2015) Targeted production of secondary metabolites by coexpression of non-ribosomal peptide synthetase and prenyltransferase genes in Aspergillus. Appl Microbiol Biotechnol 99:4213–4223
69. Kruszewska JS, Butterweck AH, Kurzatkowski W, Migdalski A, Kubicek CP, Palamarczyk G (1999) Overexpression of the Saccharomyces cerevisiae mannosylphosphodolichol synthase-encoding gene in Trichoderma reesei results in an increased level of protein secretion and abnormal cell ultrastructure. Appl Environ Microbiol 65(6):2382–2387
70. Beck PJ, Orlean P, Albright C, Robbins PW, Gething MJ, Sambrook JF (1990) The Saccharomyces cerevisiae DPM1 gene encoding dolichol-phosphate-mannose synthase is able to complement a glycosylation-defective mammalian cell line. Mol Cell Biol 10(9):4612–4622
71. Kubicek CP, Panda T, Schreferl-kunar G, Gruber F, Messner R (1987) O-linked but not N-linked glycosylation is necessary for the secretion of endoglucanases I and II by Trichoderma reesei. Can J Microbiol 33(8):698–703
72. Kruszewska JS, Palamarczyk G, Kubicek CP (1990) Stimulation of exoprotein secretion by choline and Tween 80 in Trichoderma reesei QM 9414 correlates with increased activities of dolichol phosphate mannose synthase. J Gen Microbiol 136(7):1293–1298
73. Nyyssönen E, Keränen S, Penttilä M, Demolder J, Contreras R (1995) Protein production by the filamentous fungus Trichoderma reesei: secretion of active antibody molecules. Can J Bot 73(S1):885–890
74. Mach RL, Schindler M, Kubicek CP (1994) Transformation of Trichoderma reesei based on hygromycin B resistance using homologous expression signals. Curr Genet 25(6):567–570
75. Gruber F, Visser J, Kubicek C, de Graaff L (1990) Cloning of the Trichoderma reesei pyrG gene and its use as a homologous marker for a high-frequency transformation system. Curr Genet 18(5):447–451
76. Kubicek-Pranz EM, Gruber F, Kubicek CP (1991) Transformation of Trichoderma reesei with the cellobiohydrolase II gene as a means for obtaining strains with increased cellulase production and specific activity. J Biotechnol 20(1):83–94
77. Kruszewka J, Messner R, Kubicek CP, Palamarczyk G (1989) O-Glycosylation of proteins by membrane fractions of Trichoderma reesei QM 9414. J Gen Microbiol 135(2):301–307
78. Wang W, Meng F, Liu P, Yang S, Wei D (2014) Construction of a promoter collection for genes co-expression in filamentous fungus Trichoderma reesei. J Ind Microbiol Biotechnol 41(11): 1709–1718
79. Qin Y, Wei X, Song X, Qu Y (2008) Engineering endoglucanase II from Trichoderma reesei to improve the catalytic efficiency at a higher pH optimum. J Biotechnol 135(2):190–195
80. Schulein M (1997) Enzymatic properties of cellulases from Humicola insolens. J Biotechnol 57(1-3):71–81
81. de Groot MJ, Bundock P, Hooykaas PJ, Beijersbergen AG (1998) Agrobacterium tumefaciens-mediated transformation of filamentous fungi. Nat Biotechnol 16(9):839–842

82. Miller GL (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal Chem 31(3):426–428

83. Escalante R, Vicente JJ (2000) Dictyostelium discoideum: a model system for differentiation and patterning. Int J Dev Biol 44(8):819–835

84. Stevense M, Chubb JR, Muramoto T (2011) Nuclear organization and transcriptional dynamics in Dictyostelium. Develop Growth Differ 53(4):576–586

85. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov BA, Rivero F, Bankier AT, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, Hall N, Anjard C, Hemphill L, Bason N, Farbrother P, Desany B, Just E, Morio T, Rost R, Churcher C, Cooper J, Haydock S, van Driessche N, Cronin A, Goodhead I, Muzny D, Mourier T, Pain A, Lu M, Harper D, Lindsay R, Hauser H, James K, Quiles M, Madan Babu M, Saito T, Buchrieser C, Wardroper A, Felder M, Thangavelu M, Johnson D, Knights A, Loulseged H, Mungall K, Oliver K, Price C, Quail MA, Urushihara H, Hernandez J, Rabbinowitsch E, Steffen D, Sanders M, Ma J, Kohara Y, Sharp S, Simmonds M, Spiegler S, Tivey A, Sugano S, White B, Walker D, Woodward J, Winckler T, Tanaka Y, Shaulsky G, Schleicher M, Weinstock G, Rosenthal A, Cox EC, Chisholm RL, Gibbs R, Loomis WF, Platzer M, Kay RR, Williams J, Dear PH, Noegel AA, Barrell B, Kuspa A (2005) The genome of the social amoeba Dictyostelium discoideum. Nature 435(7038):43–57

86. Urushihara H (2008) Developmental biology of the social amoeba: history, current knowledge and prospects. Develop Growth Differ 50(Suppl 1): S277–S281

87. Gaudet P, Fey P, Chisholm R (2008) Dictyostelium discoideum: the Social Ameba. CSH Protoc 2008:pdb emo109

88. Devreotes PN, Zigmond SH (1988) Chemotaxis in eukaryotic cells: a focus on leukocytes and Dictyostelium. Annu Rev Cell Biol 4:649–686

89. Jin T, Xu X, Fang J, Isik N, Yan J, Brzostowski JA, Hereld D (2009) How human leukocytes track down and destroy pathogens: lessons learned from the model organism Dictyostelium discoideum. Immunol Res 43(1-3):118–127

90. Arya R, Bhattacharya A, Saini KS (2008) Dictyostelium discoideum – a promising expression system for the production of eukaryotic proteins. FASEB J 22(12):4055–4066

91. Veltman DM, Akar G, Bosgraaf L, Van Haastert PJ (2009) A new set of small, extrachromosomal expression vectors for Dictyostelium discoideum. Plasmid 61(2):110–118

92. Williams KL, Emslie KR, Slade MB (1995) Recombinant glycoprotein production in the slime mould Dictyostelium discoideum. Curr Opin Biotechnol 6(5):538–542

93. Asgari S, Arun S, Slade MB, Marshall J, Williams KL, Wheldrake JF (2001) Expression of growth factors in Dictyostelium discoideum. J Mol Microbiol Biotechnol 3(3):491–497

94. Voith G, Dingermann T (1995) Expression of the human muscarinic receptor gene m2 in Dictyostelium discoideum. Biotechnology (N Y) 13(11): 1225–1229

95. Nguyen HN, Yang JM, Afkari Y, Park BH, Sesaki H, Devreotes PN, Iijima M (2014) Engineering ePTEN, an enhanced PTEN with increased tumor suppressor activities. Proc Natl Acad Sci U S A 111(26):E2684–E2693

96. Vervoort EB, van Ravestein A, van Peij NN, Heikoop JC, van Haastert PJ, Verheijden GF, Linskens MH (2000) Optimizing heterologous expression in dictyostelium: importance of 5′ codon adaptation. Nucleic Acids Res 28(10):2069–2074

97. Behrmann E, Muller M, Penczek PA, Mannherz HG, Manstein DJ, Raunser S (2012) Structure of the rigor actin-tropomyosin-myosin complex. Cell 150(2):327–338

98. Kon T, Oyama T, Shimo-Kon R, Imamula K, Shima T, Sutoh K, Kurisu G (2012) The 2.8 A crystal structure of the dynein motor domain. Nature 484(7394):345–350

99. Basu S, Fey P, Pandit Y, Dodson R, Kibbe WA, Chisholm RL (2013) DictyBase 2013: integrating multiple Dictyostelid species. Nucleic Acids Res 41(Database issue):D676–D683

100. Fey P, Dodson RJ, Basu S, Chisholm RL (2013) One stop shop for everything Dictyostelium: dictyBase and the Dicty Stock Center in 2012. Methods Mol Biol 983:59–92

Part IV

Higher Eukaryotic Expression Hosts

12 Fundamentals of Baculovirus Expression and Applications

Thomas A. Kost and Christopher W. Kemp

Abstract

In 1982 *E. coli* produced human insulin, the world's first recombinant DNA drug, was approved by the FDA. Since this historical event, remarkable progress has been made in developing bacterial, yeast, mammalian and insect cell protein expression systems that are used to produce recombinant proteins for both research and clinical applications. Of the available approaches, the insect cell based baculovirus expression vector system (BEVS) has proven to be a particularly adaptable system for producing a diverse collection of proteins. Along with *E. coli*, the system has been valuable for the production of proteins for structural studies, including adequate quantities of difficult to produce G protein-coupled receptors. BEVS has also been used for production of the human papilloma virus vaccine, Cervarix, the first FDA approved insect cell produced product and FluBlok, a vaccine based on the influenza virus hemagglutinin protein. Baculoviruses, modified to contain mammalian promoters (BacMam viruses), have proven to be efficient gene delivery vectors for mammalian cells and provide an alternative transient mammalian cell based protein expression approach to that of plasmid DNA based transfection methodologies. Here we provide an update on recent advances in baculovirus vector development with a focus on the numerous applications of these viruses in basic research and biotechnology.

T.A. Kost (✉)
Molecular Discovery Research, GlaxoSmithKline,
Research Triangle Park, NC 27709, USA
e-mail: thomasakost@gmail.com

C.W. Kemp
Kempbio, Inc 5119 Pegasus Court, Frederick,
MD 21704, USA

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_12

Keywords

Baculovirus expression vector system (BEVS) • Insect cells • Baculovirus display • BacMam • Gene delivery into mammalian cells

12.1 Introduction

The ability to rapidly produce a variety of functional recombinant proteins underpins many aspects of biomedical research studies. A number of recent reviews have covered the properties of many of the currently available protein production systems [1–3]. Each system has pros and cons regarding ease of use, protein production levels, cost, post-translational modification capabilities and biosafety requirements. The BEVS, possessing many advantageous features, has developed into one of the most widely used protein expression methodologies for a variety of biotechnology applications. The system is relatively easy to establish in the laboratory and the technology has evolved to the point where recombinant viruses can be readily generated, identified and quantitated. Host insect cells grow in suspension culture at 28 °C in the absence of CO_2 and serum free media formulations are commercially available. The scale-up of insect cell cultures for recombinant protein and baculovirus production can be carried out using a variety of bioreactors, from shake flasks to stirred tanks and single use wavebag systems [4–6]. The viruses do not replicate in mammalian cells, and thus have an inherently low risk biosafety profile [7, 8].

In their natural environment, outside of the research laboratory, the Baculoviridae family of viruses infect arthropods [9, 10]. The viruses take on two forms, termed occlusion derived virus (ODV) and budded virus (BV). The ODV is used as an insecticide [11–13] whereas the BV form produced by infected cultured insect cells is used in the BEVS system. The prototype baculovirus commonly used for deriving recombinant viruses is the *Autographa californica* multiple nuclear polyhedrosis virus (AcNPV). Virus particles have a distinctive enveloped rod shaped morphology with a size range of 30–60 nm in diameter and 250–300 nm in length. The genome of AcNPV is a circular double stranded DNA of approximately 134 kB and the entire DNA sequence has been determined and mapped with 156 open reading frames [14]. The initial study describing a baculovirus vector for the production of recombinant human beta interferon in insect cells was published over 30 years ago by Smith et al. [15]. Soon thereafter Pennock et al. [16] described a baculovirus vector expressing *E. coli* β-galactosidase. For a number of reasons the original vectors utilized the baculovirus polyhedrin protein gene promoter to regulate expression of the recombinant protein. Most importantly polyhedrin protein is not essential for baculovirus replication in cultured insect cells [17]. The polyhedrin promoter is highly transcribed during viral infection and the absence of the polyhedrin protein provides a means for the visual identification of recombinant virus derived plaques. These pioneering studies set the stage for the further development of this unique protein expression system [18–20]. In most cases baculovirus vectors are derived by either co-transfection of insect cells with a plasmid transfer vector carrying the gene(s) of interest together with baculovirus DNA [9, 21–25] or via the bacmid system originally developed by Luckow et al. [26]. Both approaches rely on the fact that circular baculovirus DNA is capable of initiating a complete replication cycle in transfected insect cells. With current technologies, once the appropriate transfer vector has been constructed, recombinant viruses can be isolated within a week or less. A large number of transfer vectors, baculovirus DNAs and instructional materials are available through commercial sources. Publications by Jarvis [27] and Osz-Papai [28] extensively detail the steps involved in producing recombinant proteins with the BEVS using the commonly used Sf9 insect cell host line.

The utility of baculovirus vectors for producing recombinant proteins in insect cells has been considerably extended by the capability to display proteins on the viral surface [29–31],

Table 12.1 Baculovirus/BacMam virus applications

Production of individual recombinant proteins in insect cells for:
Mechanistic and structural studies
Assay development
Use as immunogens
Multi-subunit protein complexes including virus like particles (VLPs)
Production of infectious viruses such as adeno associated virus (AAV)
Baculovirus surface display (immunization, receptor assays, imaging, virus targeting)
BacMam virus for gene delivery into mammalian cells (broad application as an alternative to transfection, electroporation and other viral based gene delivery methods)
BacMam virus for launching virus infections (AAV, lentiviruses, hepatitis B virus)
BacMam/Display virus as a potential vaccine
BacMam virus as a potential gene therapy vector
BacMam virus for interfering and MicroRNA delivery

referred to as baculo-display, and the development of modified viruses containing mammalian derived promoters, referred to as BacMam viruses, for gene delivery into mammalian cells [32–34]. As shown in Table 12.1 these developments have significantly expanded the application of baculovirus derived vectors for recombinant protein production in mammalian cells, mammalian cell based assays, gene delivery into cell types that may be difficult to transfect, vaccine development and potentially gene therapy.

12.2 BEVS

Since it inception numerous improvements have been made to recombinant baculovirus generation techniques. Nonetheless, the vast majority of vectors still exploit the polyhedrin or p10 gene promoter to control recombinant protein expression in insect cells. Also in most instances Sf9, Sf21 or High Five™ serve as host cell lines [24]. Many different assays have been developed to quantitate recombinant viruses [35]. The plaque assay is considered the gold standard; however, it is tedious and requires 5–7 days to complete [24]. A rapid quantitation method that employs a flow cytometer designed to enumerate virus particles based on their ability to simultaneously bind nucleic acid and protein specific dyes has recently been described [36].

An area of insect cell line engineering that has seen substantial progress over the past decade is glycoengineering. A number of specialized cell lines have been developed that are designed to produce recombinant proteins that have more mammalian like glycosylation patterns than those produced by unmodified insect cells [37]. A novel approach to facilitate glycoengineering that utilizes baculovirus inducible glycogene expression of engineered Sf9 cells has recently been reported [38]. This baculovirus infection mediated induction methodology, which takes advantage of the finding that the baculovirus 39 K promoter is silent in Sf9 cells in the absence of baculovirus infection, may prove extremely useful as a general approach for engineering baculovirus host cells [39].

A number of modifications to baculovirus vectors aimed at increasing the quantity and quality of baculovirus expressed recombinant proteins are discussed in a detailed review by Hitchman et al. [40]. Modifications such as the deletion of the virus chitinase and cathepsin genes have been reported to improve the stability of a number of expressed proteins. Oxford Expression Technologies markets a baculovirus vector system termed flashBAC™ULTRA that is deficient in chiA, v-cath, p10, p26 and p74. It will be interesting over time to observe the protein expression levels obtained with this vector. It has been reported that incorporation of vankyrin sequences into a baculovirus transfer plasmid can improve recombinant protein production [41]; however, no further studies have been published in the literature extending this observation. A recent report has described the development of a novel baculovirus vector expression cassette containing rearranged baculovirus-derived regulatory elements [42]. The expression cassette contains a cDNA encoding the baculovirus transactivation factors IE1 and IE0 expressed under the control of the polyhedrin promoter and a homologous repeated transcription enhancer sequence operatively cis-linked to the baculovirus p10 promoter or to chimeric promoters containing p10. A fourfold increase in recombinant protein expression was reported using this vector as compared to a standard baculovirus vector. Another interesting approach to increase protein expression using the BEVS may be to fuse the coding sequence of the protein of interest to a partial polyhedrin protein coding sequence [43]. This study reported that the production of green fluorescent protein and the E2 protein of classical swine fever virus was significantly increased following such a fusion.

The BEVS frequently serves as a biological factory for producing virus-like particles (VLPs) [44–46]. VLPs are designed to resemble viruses; however, they are non-infectious since they do not contain viral genetic material. A clinically available VLP based vaccine for the prevention of human papilloma virus infection, Cervarix, is produced using the BEVS [47]. The production of VLPs is a complex process, requiring specific amounts of viral subunit proteins produced at the appropriate time to correctly assemble into VLPs that closely resemble those formed by infection with wild type virus. A number of attributes make the BEVS attractive for producing populations of VLPs. One can attempt to control the quantity of viral capsid proteins produced in the infected insect cell by using different baculovirus promoters, varying the ratio and quantity of infecting baculoviruses, carefully controlling bioreactor conditions and defining the time of VLP harvest.

In addition to the production of VLPs the BEVS has been used successfully to produce functional adeno-associated virus (AAV) virus. First reported by Urabe et al. the process makes use of baculovirus vectors that express the necessary AAV components for the assembly of functional AAV [48]. The entire process has been scaled-up to production levels [49, 50]. In 2012, Glybera, a recombinant AAV which compensates for lipoprotein lipase deficiency and is produced using the BEVS became the first human gene therapy product to gain approval by the European Medicines Agency [51]. Recently Mietzsch et al. have reported on the development of OneBac, which consists of a panel of stable Sf9 cell lines harboring silent copies of the AAV1-12 rep and cap genes that are induced upon infection with a single baculovirus that also carries the AAV genome [52]. This approach facilitates the production of multiple AAV serotypes using the BEVS.

12.3 Baculovirus Display

The ability to display heterologous proteins on the surface of baculovirus was first described by Boublik et al. [29] and recently reviewed by Grabherr and Ernst [31]. In most instances baculovirus vectors are employed that allow fusion of the protein of interest into the baculovirus gp64 envelope protein. The expression of the fusion protein is typically under the control of a polyhedrin gene promoter. Following infection of insect cells the amplified baculovirus is coated with the wild type gp64 envelope protein together with the gp64 fusion protein. As first shown by Lindley et al. baculoviruses displaying such fusion proteins could be used as an immunogen for developing monoclonal antibodies specific for the displayed fusion protein [53]. This technology is proving to be very useful for the development of baculovirus based vaccines as reviewed by Lin et al. [54] and Paul et al. [55]. The display of protein fragments on the viral surface may also be useful for enhancing and targeting baculovirus entry into mammalian cells. Baculovirus display of a short peptide motif from gp350/220 of Epstein-Barr virus has been reported to enhance the entry of the virus into lymphocytes [56].

The expression of a functional membrane protein on the surface of baculovirus particles expressing a G-protein coupled receptor (GPCR) was initially reported by Loisel et al. [57]. The authors showed that viral particles released from Sf9 cells infected with a recombinant baculovirus coding for the human beta 2-adrenergic receptor (beta 2AR) contain glycosylated and biologically active beta 2AR. Following this observation a similar approach has been used to produce baculovirus particles coated with a variety of GPCRs

[58]. The GPCRs displayed on the virus particles have been shown to couple functionally with G protein subunits and provide a useful reagent for studying functionally constituted receptor-ligand complexes [59, 60]. These virus particles also provide a unique means of immunogen presentation for attempting to raise antibodies directed against difficult to purify membrane bound proteins. With this in mind Saitoh et al. developed a gp64 transgenic mouse line for immunization in order to reduce the development of gp64 antibodies following immunization with a baculovirus displaying membrane proteins [61]. Immunization of these mice with baculovirus particles displaying the peptide transporter PepT1 resulted in the development of a large number of monoclonal antibodies specific for the transporter. A similar approach has been taken by Ramadhanti et al. to raise a monoclonal antibody to the C-terminal region of Aquaporin-4 [62].

12.4 BacMam Virus Gene Delivery into Mammalian Cells

The finding that modified baculoviruses containing mammalian regulatory sequences, commonly referred to as BacMam viruses, could be used as gene delivery vectors for mammalian cells significantly broadened the range of baculovirus applications. The initial reports of gene delivery by BacMam viruses focused on transduction of cells of hepatic origin [32, 33]. Soon afterward it became evident that these viruses could be used to transduce a wide variety of cell types [34, 63]. This is an additive advantage to BEVS, since one is no longer limited to the use of insect cell lines for producing recombinant proteins. BacMam viruses have been used as gene delivery vectors for assay development, protein production, vaccine and gene therapy development, launching virus infections, producing viruses and basic research studies. The viruses offer many advantages for gene delivery into mammalian cells. These include ease of use, a large cloning capacity, ability to transduce a wide variety of cell types, little to no observable cell toxicity, ability to transduce with multiple viruses, a low risk biosafety profile, low cost, and a high level of reproducibility. For most applications BacMam viruses are used for transient protein expression. However, Merrihew et al. showed that transduction of CHO cells with a BacMam virus expressing a G418 resistance marker and GFP reporter gene resulted in the isolation of stable lines expressing GFP following antibiotic selection [64]. Interestingly, of the four clonal cell lines that were analyzed each had only a single integrated copy of the GFP expression cassette. Additional approaches to prolong gene expression have also been reported. These include the use of oriP/EBNA-1 [65] and incorporation of the *Sleeping Beauty* transposon into a BacMam vector [66].

The initial BacMam vectors contained a mammalian virus promoter to control expression of the recombinant protein encoding sequence and a polyadenylation signal. Over time the vectors have been enhanced by incorporating additional elements. Figure 12.1 depicts the expression cassette of a vector commonly referred to as BacMam version 2 developed by FM Boyce (unpublished results). In addition to a modified CMV promoter containing an intron the version 2 vector contains a copy of the Woodchuck post-translational regulatory element (WPRE), which has been shown to enhance BacMam virus-mediated gene expression in mammalian cells [67]. It also contains a copy of the vesicular stomatitis glycoprotein G (VSV-G) gene which has been shown to enhance viral transduction [68]. In this case VSV-G expression is controlled by a polyhedrin gene promoter. Thus, the BacMam virus produced by infected insect cells is coated with both gp64 and VSV-G; however, VSV-G is not produced by transduced mammalian cells since the polyhedrin gene promoter is inactive in mammalian cells. Although the CMV promoter is typically used in BacMam vectors other promoters, such as CAG [63], Rous sarcoma virus (RSV) [33], CBA, EF1-α [69] and WSSV ie1 [70] have been used. The ability to incorporate a variety of beneficial regulatory elements into BacMam vectors is a significant advantage of this technology. For an overview of some of the aspects one should consider when transducing

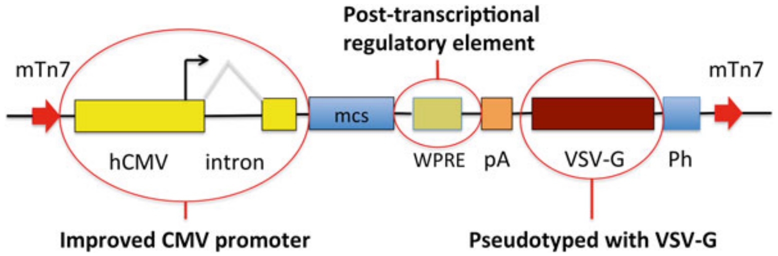


Fig. 12.1 Diagram of BacMam virus version 2.0 with enhanced expression features. The diagram shows the important mammalian cell components engineered into the BacMam virus shuttle vector commonly referred to as version 2.0. These include (*1*) An improved human CMV promoter containing an intron (*2*) woodchuck post-transcriptional regulatory element (WPRE) and (*3*) vesicular stomatitis virus-G protein (VSV-G), which is expressed off the viral polyhedrin promoter and thus is not expressed in mammalian cells. These elements serve to enhance the transduction efficiency of the derived BacMam viruses by broadening the host cell range and enhancing protein expression levels. (*mTn7* miniTn7, *mcs* multiple cloning site, *pA* polyadenylation signal)

mammalian cells with BacMam vectors see the articles by Airenne [71] and Sung et al. [72].

A number of reviews have been published on the use of BacMam viruses for cell assay development [73–76]. A luciferase enzyme fragment complementation assay to identify nuclear-factor-e2-related transcription factor 2 activators described by Xie et al. provides a good example of the capability of BacMam virus transduction to deliver two protein components requiring a close interaction to provide a functional enzymatic readout [77]. An assay designed to identify modulators of human epithelial sodium channels (ENaCs) employed a BacMam virus to transiently express the ENaC alpha subunit in a stable HEK293 cell line expressing the ENaC beta and gamma subunit variants. In this instance BacMam virus delivery provided a means to reduce the cellular toxicity associated with long-term expression of the ENaC alpha subunit [78]. Mazina et al. have recently described the delivery of cAMP based biosensors via BacMam viruses [79]. The capability of BacMam viruses to successfully transduce stem cells has been reviewed by Sung et al. [72]. This attribute of BacMam viruses provides a powerful methodology for modifying stem cells and studying their biology.

BacMam virus transduction has reached the point where transduced cells can be used to produce large quantities of recombinant proteins. Scott et al. reported the efficient expression of secreted proteases by BacMam virus transduced HEK293 cells [80]. The production of a large number of proteases was examined with 14 of 16 proteases produced at 10–30 mg or more of purified protein per liter of culture medium. Recently Goehring et al. have published protocols for the large-scale expression of membrane proteins for structural studies using BacMam virus to transduce a modified HEK 293 cell line [81]. Purified chicken acid sensing-ion channel 1a and *Caenorhabditis elegans* glutamate-gated chloride channel were produced for x-ray crystallography. As shown in Fig. 12.2 BacMam virus transduction has also been used successfully for the transient production of high levels of recombinant IgG. In this instance a yield of 140 mg/L of purified antibody was obtained from the transduced HEK293F cells. As discussed previously the BEVS has been used frequently for the production of VLPs. As shown in Fig. 12.3 initial studies indicate that BacMam transduction of mammalian cells can also be used for the production of VLPs. An advantage of this approach versus using the BEVS is that BacMam virus replication does not occur in the transduced HEK 293 cells, thus facilitating the purification of VLPs. Lesch et al. have also shown that BacMam transduction can be used to produce functional lentiviral vectors [82]. These examples clearly establish the exciting potential of BacMam virus transduction of mammalian cells as an alternative approach to BEVS for recombinant protein production.

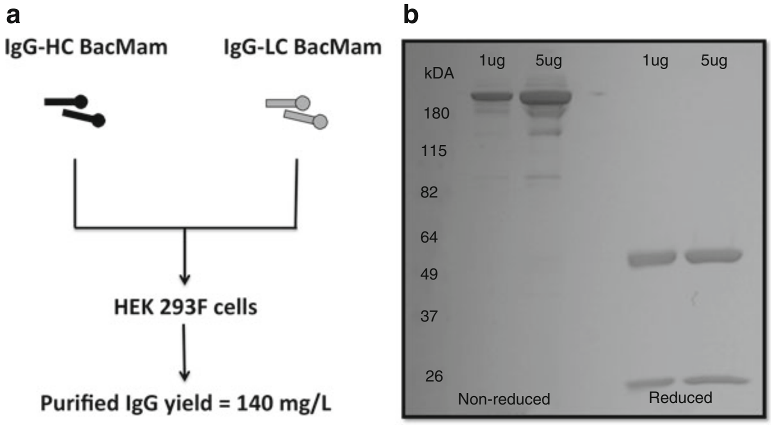


Fig. 12.2 Production of a humanized rIgG by BacMam virus transduced HEK293F cells. (**a**) BacMam viruses containing heavy and light IgG coding sequences were used to transduce HEK293F cells cultured at 37 °C in serum-free Freestyle medium in a 10-L stirred tank reactor. The multiplicity of transduction ratio was 25:25 for the HC and LC viruses, respectively. The culture supernatant was harvested at 72 h post transduction and analyzed by gel electrophoresis. (**b**) Coomassie stained SDS PAGE gel of rIgG produced following BacMam virus transduction. Gels were run under non-reduced and reducing conditions. The total yield of purified rIgG was 140 mg/L. (HEK293F cells and Freestyle medium were obtained from Life Technologies)

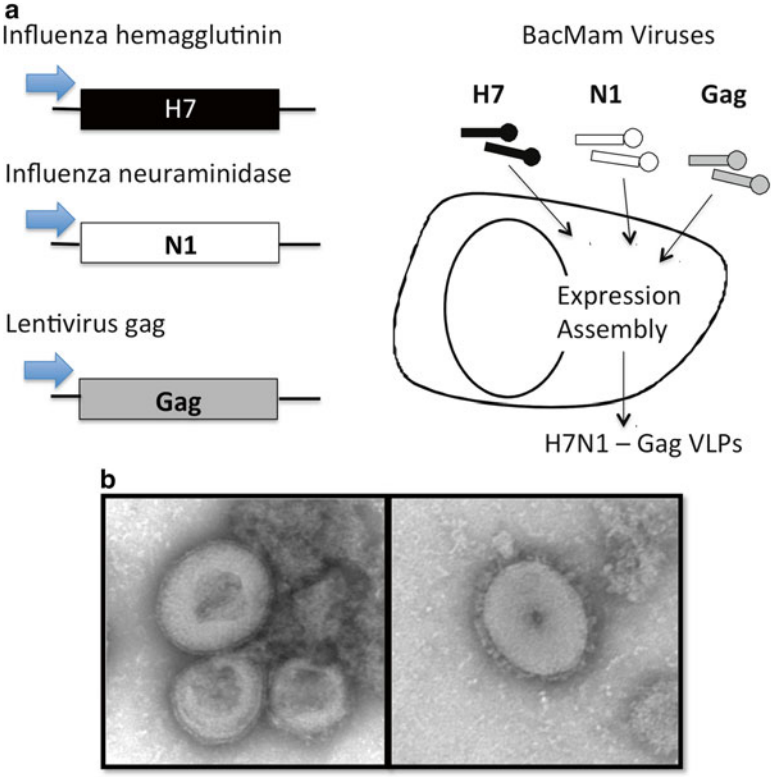


Fig. 12.3 Production of hybrid H7N1 influenza VLPs in HEK293F cells using BacMam virus transduction. (**a**) Expression constructs for H7, N1, and Gag are shown. The CMV promoter is indicated with *blue arrow*. The transduction process for H7N1 VLP production is depicted on the right. HEK 293 F cells were cultured in a 1 L shake flask in serum-free Freestyle medium at 37 °C and transduced at a multiplicity of transduction of 14:14:7 with the H7, N1 and Gag expressing viruses, respectively. Culture supernatant was harvested at 120 h post transduction and VLPs were purified by centrifugation. (**b**) Transmission electron micrographs of purified VLPs (Courtesy of Peter Pushko, Medigen, Inc.)

A considerable amount of exciting research is also being conducted toward utilizing either BEVS or BacMam viruses for vaccine and gene therapy applications. Upon writing this article a search of the PubMed database using the terms "baculovirus vaccines" brings up 853 entries and "baculovirus gene therapy" 374 entries. It is beyond the scope of this article to go into details on these studies. For overviews of these areas a number of recent reviews have focused on vaccines [44, 83–85] and gene therapy applications [86–90].

12.5 Conclusion and Future Perspectives

Recombinant baculoviruses provide a powerful tool for a wide variety of biotechnology applications. These range from producing recombinant proteins in insect cells to baculo-display and gene delivery into mammalian cells. To date no other viral vector system has been described that can be used for such a wide repertoire of applications. One can unquestionably envision continued discoveries and improvements in vector design and host cell engineering that will further enhance the capability of this unique expression system for both insect and mammalian cell applications. For example, one could engineer a CHO or HEK293 host cell line with features intended to enhance BacMam transduction efficiency and increase transient protein production levels. The ability to construct hybrid viruses by incorporating regulatory elements from mammalian viruses also provides the potential to enhance BacMam virus transduction [90]. Future animal model studies may provide knowledge that leads to creative approaches for enhancing the combination of baculo-display and BacMam transduction to increase the vector's potential as an immunogen that may lead to its eventual use as an animal or human vaccine candidate. A topic that deserves further attention is the development of a baculovirus reference standard [91]. As more baculovirus applications move toward the clinic it will be important to have a baculovirus reference material available for reliable quantitation of virus preparations within and between laboratories and manufacturing facilities.

Acknowledgements The authors are grateful to Frederick R. Boyce, Massachusetts General Hospital for providing the schematic of BacMam version 2.0 shown in Fig. 12.1 and Peter Pushko of Medigen, Inc. for providing the information shown in Fig. 12.3.

References

1. Assenberg R, Wan PT, Geisse S, Mayr LM (2013) Advances in recombinant protein expression for use in pharmaceutical research. Curr Opin Struct Biol 23:393–402
2. Bandaranayake AD, Almo SC (2014) Recent advances in mammalian protein production. FEBS Lett 588:253–260
3. Cuozzo JW, Soutter HH (2014) Overview of recent progress in protein-expression technologies for small-molecule screening. J Biomol Screen 19:1000–1013
4. Drugmand JC, Schneider YJ, Agathos SN (2012) Insect cells as factories for biomanufacturing. Biotechnol Adv 30:1140–1157
5. Kadwell SH, Hardwicke PI (2007) Production of baculovirus expressed recombinant proteins in wave bioreactors. Methods Mol Biol 388:247–266
6. Aucoin MG, Mena JA, Kamen AA (2010) Bioprocessing of baculovirus vectors: a review. Curr Gene Ther 10:174–186
7. Kost TA, Condreay JP (2002) Innovations – biotechnology: baculovirus vectors as gene transfer vectors for mammalian cells: biosafety considerations. Appl Biosaf 7:167–169
8. Kost TA, Condreay PJ, Mickelson CA (2006) Biosafety and viral gene transfer vectors. In: Fleming DO, Hunt DL (eds) Biological safety, principles and practices, 4th edn. American Society for Microbiology Press, Washington D.C, pp 509–530
9. O'Reilly DR, Miller LK, Luckow VA (1994) Baculovirus expression vectors. Oxford University Press, New York
10. Rohrmann GF (2013) Baculovirus molecular biology, 3rd edn. National Center for Biotechnology Information (US), Bethesda. http://www.ncbinlm.nih.gov/books/NBK114593/
11. Inceoglu AB, Kamita SG, Hammock BD (2006) Genetically modified baculoviruses: a historical overview and future outlook. Adv Virus Res 68:323–360
12. Szewczyk B, Hoyos-Carvajal L, Paluszek M, Skrzecz I, Lobo de Souza M (2006) Baculoviruses – re-emerging biopesticides. Biotechnol Adv 24:143–160
13. Shim HJ, Choi JY, Wang Y, Tao XY, Liu Q, Roh JY, Kim JS, Kim WJ, Woo SD, Jin BR, Je YH (2013) NeuroBactrus, a novel, highly effective, and environmentally friendly recombinant baculovirus insecticide. Appl Environ Microbiol 79:141–149

14. Ayres MD, Howard SC, Kuzio J, Lopez-Ferber M, Possee RD (1994) The complete DNA sequence of Autographa californica nuclear polyhedrosis virus. Virology 202:586–605
15. Smith GE, Summers MD, Fraser MJ (1983) Production of human beta interferon in insect cells infected with a baculovirus expression vector. Mol Cell Biol 3:2156–2165
16. Pennock GD, Shoemaker C, Miller LK (1984) Strong and regulated expression of Escherichia coli beta-galactosidase in insect cells with a baculovirus vector. Mol Cell Biol 4:399–406
17. Smith GE, Fraser MJ, Summers M (1983) Molecular engineering of the Autographa californica nuclear polyhedrosis virus genome: deletion mutations within the polyhedron gene. J Virol 46:584–593
18. Summers MD (2006) Milestones leading to the genetic engineering of baculoviruses as expression vector systems and viral pesticides. Adv Virus Res 68:3–73
19. Kost TA, Condreay JP, Jarvis DL (2005) Baculovirus as versatile vectors for protein expression in insect and mammalian cells. Nat Biotechnol 23:567–575
20. Van Oers MM, Pijlman GP, Vlak JM (2015) Thirty years of baculovirus-insect cell protein expression: From dark horse to mainstream technology. J Gen Virol 96:6–23
21. Kitts PA, Possee RD (1993) A method for producing recombinant baculovirus expression vectors at high frequency. Biotechniques 14:810–817
22. Murhammer DW (ed) (2007) Baculovirus and insect cell expression protocols. Springer, New York, Methods Mol Biol
23. Possee RD, Hitchman RB, Richards KS, Mann SG, Siaterli E, Nixon CP, Irving H, Assenberg R, Alderton D, Owens RJ, King LA (2008) Generation of baculovirus vectors for the high-throughput production of proteins in insect cells. Biotechnol Bioeng 101:1115–1122
24. Jarvis DL (2009) Baculovirus-insect cell expressions systems. Methods Enzymol 463:191–221
25. Cremer H, Bechtold I, Mahnke M, Assenberg R (2014) Efficient processes for protein expression using recombinant baculovirus particles. Methods Mol Biol 1104:395–417
26. Luckow VA, Lee SC, Barry GF, Olins PO (1993) Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of froreign genes into a baculovirus genome propagated in *Escherichia coli*. J Virol 67:4566–4579
27. Jarvis DL (2014) Recombinant protein expression in baculovirus-infected insect cells. Methods Enzymol 536:149–163
28. Osz-Papai J, Radu L, Abdulrahman W, Kolb-Cheynel I, Troffer-Charlier N, Birck C, Poterszman A (2015) Insect cells-baculovirus system for the production of difficult to express proteins. Methods Mol Biol 1258:181–205
29. Boublik Y, DiBonito P, Jones IM (1995) Eukaryotic virus display: engineering the major surface glycoprotein of the *Autographa californica nuclear* polyhedrosis virus (AcNPV) for the presentation of foreign protein on the virus surface. Bio/Technology 13:1079–1084
30. Grabherr R, Ernst W, Doblhoff-Dier O, Sara M, Katinger H (1997) Expression of foreign proteins on the surface of *Autographa californica* nuclear polyhedrosis virus. Biotechniques 22:730–735
31. Grabherr R, Ernst W (2013) Baculovirus for eukaryotic protein display. Curr Gene Ther 10:195–200
32. Hoffmann C, Sandig V, Jennings G, Rudolph M, Schlag P, Strauss M (1995) Efficient gene transfer into human hepatocytes by baculovirus vectors. Proc Natl Acad Sci U S A 92:10099–10103
33. Boyce FM, Bucher NL (1996) Baculovirus-mediated gene transfer into mammalian cells. Proc Natl Acad Sci U S A 93:2348–2352
34. Condreay JP, Witherspoon SM, Clay WC, Kost TA (1999) Transient and stable gene expression in mammalian cells transduced with a recombinant baculovirus vector. Proc Natl Acad Sci U S A 96:127–132
35. Roldão A, Oliveira R, Carrondo MJ, Alves PM (2009) Error assessment in recombinant baculovirus titration: evaluation of different methods. J Virol Methods 159:69–80
36. Birch A, Allen H, Kennefick K, Gugel A, Kemp CW (2014) Rapid and effective monitoring of baculovirus concentrations in bioprocess fluid using the ViroCyt® Virus Counter®. Bioprocess J 13:32–39
37. Mabashi-Asazuma H, Shi X, Geisler C, Kuo CW, Khoo KH, Jarvis DL (2013) Impact of a human CMP-sialic acid transporter on recombinant glycoprotein sialylation in glycoengineered insect cells. Glycobiology 23:199–210
38. Toth AM, Kuo CW, Khoo KH, Jarvis DL (2014) A new insect cell glycoengineering approach provides baculovirus-inducible glycogene expression and increases human-type glycosylation efficiency. J Biotechnol 182–183:19–29
39. Lin CH, Jarvis DL (2013) Utility of temporally distinct baculovirus promoters for constitutive and baculovirus-inducible transgene expression in transformed insect cells. J Biotechnol 165:11–17
40. Hitchman RB, Locanto E, Possee RD, King LA (2011) Optimizing the baculovirus expression vector system. Methods 55:52–57
41. Fath-Goodin A, Kroemer J, Martin S, Reeves K, Webb BA (2006) Polydnavirus genes that enhance the baculovirus expression vector system. Adv Virus Res 68:75–90
42. Gómez-Sebastián S, López-Vidal J, Escribano JM (2014) Significant productivity improvement of the baculovirus expression vector system by engineering a novel expression cassette. PLoS ONE 9:e96562
43. Bae SM, Kim HJ, Lee JB, Choi JB, Shin TY, Koo HN, Choi JY, Lee KS, Je YH, Jin BR, Yoo SS, Woo SD (2013) Hyper-enhanced production of foreign recom-

binant protein by fusion with the partial polyhedrin of nucleopolyhedrovirus. PLoS ONE 8:e60835

44. Mena JA, Kamen AA (2011) Insect cell technology is a versatile and robust vaccine manufacturing platform. Expert Rev Vaccines 10:1063–1081
45. Fernandes F, Teixeira AP, Carinhas N, Carrondo MJ, Alves PM (2013) Insect cells as a production platform of complex virus-like particles. Expert Rev Vaccines 12:225–236
46. Yamaji H (2014) Suitability and perspectives on using recombinant insect cells for the production of virus-like particles. Appl Microbiol Biotechnol 98:1963–1970
47. Deschuyteneer M, Elouahabi A, Plainchamp D, Plisnier M, Soete D, Corazza Y, Lockman L, Giannini S, Deschamps M (2010) Molecular and structural characterization of the L1 virus-like particles that are used as vaccine antigens in Cervarix™, the AS04-adjuvanted HPV-16 and -18 cervical cancer vaccine. Hum Vaccin 6:407–419
48. Urabe M, Ding C, Kotin RM (2002) Insect cells as a factory to produce adeno-associated virus type 2 vectors. Hum Gene Ther 13:1935–1943
49. Mena JA, Aucoin MG, Montes J, Chahal PS, Kamen AA (2010) Improving adeno-associated vector yield in high density insect cell cultures. J Gene Med 12:157–167
50. Kotin RM (2011) Large-scale recombinant adeno-associated virus production. Hum Mol Genet 20:R2–R6
51. Ylä-Herttuala S (2012) Endgame: glybera finally recommended for approval as the first gene therapy drug in the European union. Mol Ther 20:1831–1832
52. Mietzsch M, Grasse S, Zurawski C, Weger S, Bennett A, Agbandje-McKenna M, Muzyczka N, Zolotukhin S, Heilbronn R (2014) OneBac: platform for scalable and high-titer production of adeno-associated virus serotype 1–12 vectors for gene therapy. Hum Gene Ther 25:212–222
53. Lindley KM, Su JL, Hodges PK, Wisely GB, Bledsoe RK, Condreay JP, Winegar DA, Hutchins JT, Kost TA (2000) Production of monoclonal antibodies using recombinant baculovirus displaying gp-64 fusion proteins. J Immunol Methods 34:123–135
54. Lin SY, Chung YC, Hu YC (2014) Update on baculovirus as an expression and/or delivery vehicle for vaccine antigens. Expert Rev Vaccines 13:1501–1521
55. Paul A, Hasan A, Rodes L, Sangaralingam M, Prakash S (2014) Bioengineered baculoviruses as new class of therapeutics using micro and nanotechnologies: principles, prospects and challenges. Adv Drug Deliv Rev 71:115–130
56. Ge J, Huang Y, Hu X, Zhong J (2007) A surface-modified baculovirus vector with improved gene delivery to B-lymphocytic cells. J Biotechnol 129:367–372
57. Loisel TP, Ansanay H, St-Onge S, Gay B, Boulanger P, Strosberg AD, Marullo S, Bouvier M (1997) Recovery of homogeneous and functional beta 2-adrenergic receptors from extracellular baculovirus particles. Nat Biotechnol 15:1300–1304
58. Hamakubo T, Kusano-Arai O, Iwanari H (2014) Generation of antibodies against membrane proteins. Biochim Biophys Acta 1844:1920–1924
59. Mitsui K, Sakihama T, Takahashi K, Masuda K, Fukuda R, Hamana H, Sato T, Hamakubo T (2012) Functional reconstitution of olfactory receptor complex on baculovirus. Chem Senses 37:837–847
60. Veiksina S, Kopanchuk S, Rinken A (2014) Budded baculoviruses as a tool for a homogeneous fluorescence anisotropy-based assay of ligand binding to G protein-coupled receptors: the case of melanocortin 4 receptors. Biochim Biophys Acta 1838:372–381
61. Saitoh R, Ohtomo T, Yamada Y, Kamada N, Nezu J, Kimura N, Funahashi S, Furugaki K, Yoshino T, Kawase Y, Kato A, Ueda O, Jishage K, Suzuki M, Fukuda R, Arai M, Iwanari H, Takahashi K, Sakihama T, Ohizumi I, Kodama T, Tsuchiya M, Hamakubo T (2007) Viral envelope protein gp64 transgenic mouse facilitates the generation of monoclonal antibodies against exogenous membrane proteins displayed on baculovirus. J Immunol Methods 322:104–117
62. Ramadhanti J, Huang P, Kusano-Arai O, Iwanari H, Sakihama T, Misu T, Fujihara K, Hamakubo T, Yasui M, Abe Y (2013) A novel monoclonal antibody against the C-terminal region of aquaporin-4. Monoclon Antib Immunodiagn Immunother 32:270–276
63. Shoji I, Aizaki H, Tani H, Ishii K, Chiba T, Saito I, Miyamura T, Matsuura Y (1997) Efficient gene transfer into various mammalian cells, including non-hepatic cells, by baculovirus vectors. J Gen Virol 78:2657–2664
64. Merrihew RV, Clay WC, Condreay JP, Witherspoon SM, Dallas WS, Kost TA (2001) Chromosomal integration of transduced recombinant baculovirus DNA in mammalian cells. J Virol 75:903–909
65. Shan L, Wang L, Yin J, Zhong P, Zhong J (2006) An OriP/EBNA-1-based baculovirus vector with prolonged and enhanced transgene expression. J Gene Med 8:1400–1406
66. Luo WY, Shih YS, Hung CL, Lo KW, Chiang CS, Lo WH, Huang SF, Wang SC, Yu CF, Chien CH, Hu YC (2012) Development of the hybrid Sleeping Beauty: baculovirus vector for sustained gene expression and cancer therapy. Gene Ther 19:844–851
67. Mähönen AJ, Airenne KJ, Purola S, Peltomaa E, Kaikkonen MU, Riekkinen MS, Heikura T, Kinnunen K, Roschier MM, Wirth T, Ylä-Herttuala S (2007) Post-transcriptional regulatory element boosts baculovirus-mediated gene expression in vertebrate cells. J Biotechnol 131:1–8
68. Barsoum J, Brown R, McKee M, Boyce FM (1997) Efficient transduction of mammalian cells by a recombinant baculovirus having the vesicular stomatitis virus G glycoprotein. Hum Gene Ther 8:2011–2018
69. Ge J, Jin L, Tang X, Gao D, An Q, Ping W (2014) Optimization of eGFP expression using a modified baculovirus expression system. J Biotechnol 173:41–46
70. Gao H, Wang Y, Li N, Peng WP, Sun Y, Tong GZ, Qiu HJ (2007) Efficient gene delivery into mammalian

cells mediated by a recombinant baculovirus containing a whispovirus ie1 promoter, a novel shuttle promoter between insect cells and mammalian cells. J Biotechnol 13:138–143

71. Airenne K (2009) Optimization of baculovirus-mediated gene delivery into vertebrate cells. Bio Process J 8:54–59
72. Sung LY, Chen CL, Lin SY, Li KC, Yeh CL, Chen GY, Lin CY, Hu YC (2014) Efficient gene delivery into cell lines and stem cells using baculovirus. Nat Protoc 9:1882–1899
73. Boudjelal M, Mason SJ, Katso RM, Fleming JM, Parham JH, Condreay JP, Merrihew RV, Cairns WJ (2005) The application of BacMam technology in nuclear receptor drug discovery. Biotechnol Annu Rev 11:101–125
74. Davenport EA, Nuthulaganti P, Ames RS (2009) BacMam: versatile gene delivery technology for GPCR assays. Methods Mol Biol 552:199–211
75. Ames RS, Lu Q (2009) Viral-mediated gene delivery for cell-based assays in drug discovery. Expert Opin Drug Discov 4:243–256
76. Kost TA, Condreay JP, Ames RS (2010) Baculovirus gene delivery: a flexible assay development tool. Curr Gene Ther 10:168–173
77. Xie W, Pao C, Graham T, Dul E, Lu Q, Sweitzer TD, Ames RS, Li H (2012) Development of a cell-based high throughput luciferase enzyme fragment complementation assay to identify nuclear-factor-e2-related transcription factor 2 activators. Assay Drug Dev Technol 10:514–524
78. Chen MX, Gatfield K, Ward E, Downie D, Sneddon HF, Walsh S, Powell AJ, Laine D, Carr M, Trezise D (2015) Validation and optimization of novel high-throughput assays for human epithelial sodium channels. J Biomol Screen 20:242–253
79. Mazina O, Allikalt A, Heinloo A, Reinart-Okugbeni R, Kopanchuk S, Rinken A (2015) cAMP assay for GPCR ligand characterization: application of BacMam expression system. Methods Mol Biol 1272:65–77
80. Scott MJ, Modha SS, Rhodes AD, Broadway NM, Hardwicke PI, Zhao HJ, Kennedy-Wilson KM, Sweitzer SM, Martin SL (2007) Efficient expression of secreted proteases via recombinant BacMam virus. Protein Expr Purif 52:104–116
81. Goehring A, Lee CH, Wang KH, Michel JC, Claxton DP, Baconguis I, Althoff T, Fischer S, Garcia KC, Gouaux E (2014) Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. Nat Protoc 9:2574–2585
82. Lesch HP, Laitinen A, Peixoto C, Vicente T, Makkonen KE, Laitinen L, Pikkarainen JT, Samaranayake H, Alves PM, Carrondo MJ, Ylä-Herttuala S, Airenne KJ (2011) Production and purification of lentiviral vectors generated in 293 T suspension cells with baculoviral vectors. Gene Ther 18:531–538
83. Madhan S, Prabakaran M, Kwang J (2010) Baculovirus as vaccine vectors. Curr Gene Ther 10:201–213
84. Cox MM (2012) Recombinant protein vaccines produced in insect cells. Vaccine 30:1759–1766
85. Lu HY, Chen YH, Liu HJ (2012) Baculovirus as a vaccine vector. Bioengineered 3:271–274
86. Wang S, Balasundaram G (2010) Potential cancer gene therapy by baculoviral transduction. Curr Gene Ther 10:214–225
87. Rivera-Gonzalez GC, Swift SL, Dussupt V, Georgopoulos LJ, Maitland NJ (2011) Baculoviruses as gene therapy vectors for human prostate cancer. J Invertebr Pathol 107(Suppl):S59–S70
88. Lesch HP, Makkonen KE, Laitinen A, Määttä AM, Närvänen O, Airenne KJ, Ylä-Herttuala S (2011) Requirements for baculoviruses for clinical gene therapy applications. J Invertebr Pathol 107(Suppl):S106–S112
89. Airenne KJ, Hu YC, Kost TA, Smith RH, Kotin RM, Ono C, Matsuura Y, Wang S, Ylä-Herttuala S (2013) Baculovirus: an insect-derived vector for diverse gene transfer applications. Mol Ther 21:739–749
90. Zhang W, Hagedorn C, Schulz E, Lipps HJ, Ehrhardt A (2014) Viral hybrid-vectors for delivery of autonomous replicons. Curr Gene Ther 14:10–23
91. Kamen AA, Aucoin MG, Merten OW, Alves P, Hashimoto Y, Airenne K, Hu YC, Mezzina M, van Oers MM (2011) An initiative to manufacture and characterize baculovirus reference material. J Invertebr Pathol 107(Suppl):S113–S117

13 The MultiBac Baculovirus/Insect Cell Expression Vector System for Producing Complex Protein Biologics

Duygu Sari, Kapil Gupta,
Deepak Balaji Thimiri Govinda Raj, Alice Aubert,
Petra Drncová, Frederic Garzoni,
Daniel Fitzgerald, and Imre Berger

Abstract

Multiprotein complexes regulate most if not all cellular functions. Elucidating the structure and function of these complex cellular machines is essential for understanding biology. Moreover, multiprotein complexes by themselves constitute powerful reagents as biologics for the prevention and treatment of human diseases. Recombinant production by the baculovirus/insect cell expression system is particularly useful for expressing proteins of eukaryotic origin and their complexes. MultiBac, an advanced baculovirus/insect cell system, has been widely adopted in the last decade to produce multiprotein complexes with many subunits that were hitherto inaccessible, for academic and industrial research and development. The MultiBac system, its development and numerous applications are presented. Future opportunities for utilizing MultiBac to catalyze discovery are outlined.

Keywords

Protein complexes • Baculovirus • Insect cell expression system • MultiBac • Structural biology • X-ray crystallography • Electron microscopy • Virus-like particles (VLPs) • Vaccines • Gene transfer • Lepidoptera • Synthetic biology

D. Sari • K. Gupta • D.B.T.G. Raj • A. Aubert
P. Drncová • F. Garzoni
European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Unit of Virus Host-Cell Interactions, University Grenoble Alpes-EMBL-CNRS, UMI 3265, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

D. Fitzgerald
Geneva Biotech SARL,
Avenue de la Roseraie 64, 1205 Genève, Switzerland

I. Berger (✉)
European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

Unit of Virus Host-Cell Interactions, University Grenoble Alpes-EMBL-CNRS, UMI 3265, 71 avenue des Martyrs, 38042 Grenoble Cedex 9, France

School of Biochemistry, University of Bristol, Bristol BS8 1TD, UK
e-mail: iberger@embl.fr

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*, Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_13

13.1 Introduction: The Baculovirus Expression Vector System (BEVS)

More than 30 years ago, the high level production of a heterologous protein by using an insect specific baculovirus, derived from the *Autographa californica multiple nuclear polyhedrosis virus* (AcMNPV) was reported. Max Summers and coworkers produced functional human IFN-β in insect cells infected by a recombinant baculovirus [1]. This development was made possible by the previous observations that late in its viral life cycle, baculoviruses express at very high levels a protein, polyhedrin, which is not essential in laboratory culture. Substitution of the *polyhedrin* gene in the baculoviral polh locus by a foreign gene of interest resulted in comparably high-level expression of the desired gene product, driven by the polh promoter, without compromising virus infectivity and the viral life cycle [2]. Shortly after, a second study by Lois Miller and colleagues demonstrated that another very late promoter p10, showed similar characteristics and could also be used for high-level production of heterologous proteins [3].

These two seminal studies established the baculovirus/insect cell expression system as a powerful means to produce proteins recombinantly. In the three decades since these hallmark contributions, baculovirus expression has become a widely adopted technology for academic and industrial applications, in research and development as well as manufacturing, and a wide range of proteins have been made by baculovirus expression vector systems (BEVS) [2, 4–7]. Multicomponent virus-like particles (VLPs) resembling complex virus shells have been produced successfully with BEVS, including VLPs from bluetongue, rotavirus and others [8–11]. More recently, the first baculovirus produced proteins have been approved in the therapy or prevention of human disease, including vaccines against influenza (Flublok®) and cervical cancer (Cervarix®), and immune-therapeutics against tumors of the prostate (Provenge®) [4]. Moreover, baculovirus itself has emerged as a versatile tool for gene therapy, either as a production system for recombinant adeno-associated viruses [12–14] or as a DNA-based gene delivery vehicle in its own right [15, 16].

The development of BEVS for heterologous protein production and its manifold exploits has been authoritatively reviewed recently in a number of contributions, comprehensively recapitulating the technical aspects of this technology [2, 4, 5]. The subject of this present contribution is MultiBac, a particular baculovirus expression vector system developed and implemented more recently [17–25]. MultiBac was originally conceived to meet the imposing challenge of producing eukaryotic multiprotein complexes, vital cornerstones of biological activity, in high quality and quantity for high-resolution structural and functional analysis. The system has been uniquely successful in catalyzing multiprotein complex research globally. MultiBac, its ongoing development, its numerous applications and future prospects are reviewed in the following.

13.2 The MultiBac System for Expressing Eukaryotic Multiprotein Complexes

Protein complexes catalyze key functions in the cell, and as a consequence, are an intense focus of contemporary biological research efforts. Genomics and proteomics studies have underpinned that most if not all proteins in eukaryotic cells are part of larger assemblies, which in humans often comprise ten or more individual subunits. The complex interplay of proteins in these complexes is essential for cell homeostasis, biological activity and development. High-resolution functional and structural characterization of the large number of multiprotein assemblies in the cell is critical to understanding cell biology [20, 26, 27].

Multisubunit complexes may be purified from their native cell environment and their structure and function analyzed successfully at near-atomic resolution, provided they are sufficiently abundant and homogeneous. Well-known examples include RNA polymerases and ribosomes [28–31]. The overwhelming majority of multiprotein complexes in the cell, however, are characterized by low or very low abundance, which considerably complicates or even rules out their purification from native source material. Furthermore, it is becoming increasingly clear that many proteins may exist not only in one, but a number of distinct complexes, carrying out diverse functions depending on the partner molecules they associate with at a given time. Together, this often obstructs obtaining compositionally pure and homogeneous material by classical fractionation of cells and subsequent biochemical purification, notwithstanding significant progress notably in endogenous tagging methods to genomically modify endogenous proteins by powerful extraction aids such as tandem affinity purification (TAP) tags, for instance [32–34]. A solution to these issues is recombinant overproduction, enabled by the development and implementation of powerful overexpression technologies that can achieve high-level production of homogeneous and active eukaryotic complexes for detailed mechanistic analysis at the molecular level.

Recombinant protein overproduction had a profound and game-changing impact on protein science, making previously inaccessible targets readily available. A very large number of proteins, their mutants and variants have been produced recombinantly, and their structure and function determined at high resolution. The availability of entire genomes has made it possible to address the protein repertoire of cells on a system-wide scale, applying high-throughput technologies [35]. Recombinant protein expression in *E. coli* as a prokaryotic expression host has become prevalent in molecular biology laboratories world-wide. The recombinant production of protein complexes of eukaryotic origin, however, poses a number of challenges which frequently rule out prokaryotic expression hosts. Eukaryotic proteins are often large and can exceed the size range *E. coli* can overproduce efficiently (typically up to ~100 kDa). Post-translational modifications and processing are commonplace in eukaryotic proteins and can be essential for activity, but are generally not supported by a prokaryotic host. The eukaryotic protein folding machinery differs significantly from the chaperone system in *E. coli*, further restricting its utility for eukaryotic protein production. Much effort has been and is being devoted to improving prokaryotic host systems for heterologous production [36–38]. However, in many cases eukaryotic proteins and their complexes will likely require a eukaryotic expression host system for their overproduction, and if a eukaryotic system can be applied with comparable ease as *E. coli* based expression, then this system will likely be a preferred choice. The MultiBac system has been developed precisely with this intention to put in place such a eukaryotic expression system that supports high-level and high-quality production of eukaryotic proteins and their complexes, by using standard operating protocols (SOPs) which make its application comparably facile and routine as *E. coli*-based expression [18, 39–41].

13.2.1 MultiBac Developments

The baculovirus/insect cell expression system is particularly well-suited for the production of eukaryotic proteins. At the core of this expression technology is a recombinant baculovirus into which the heterologous genes of interest have been inserted. This composite baculovirus is then used to infect insect cell cultures grown in the laboratory. MultiBac is a more recent baculovirus/insect cell system which has been specifically tailored for the overproduction of eukaryotic complexes that contain many subunits [40]. An important prerequisite for the efficient expression of eukaryotic proteins and their complexes is easy-to-use reagents for (multi)gene assembly and delivery. Equally required are robust and standardized protocols for all steps involved in the expression experiment, from gene to purified protein complex. These steps should ideally be implemented as standard operating procedures (SOPs), especially in laboratories where the expression experiment itself and its optimizations are not the primary objective, but rather the protein complex and the determination of its structure and mechanism within a reasonable time-frame. The implementation of such SOPs will then enable non-specialist users to apply the technology with relative ease. The MultiBac system has been designed to meet these requirements [18, 40, 41].

MultiBac consists of an engineered baculovirus that has been optimized for multiprotein complex expression [17] (Fig. 13.1). The MultiBac baculovirus exists as a bacterial artificial chromosome (BAC) in *E. coli* cells (DH10MultiBac or DH10MB). The replicon (F-factor) present on the BAC restricts its copy number to (typically) one [42]. The MultiBac genome has been modified by deleting proteolytic and apoptotic functionalities from the baculoviral genome that were found to be detrimental for the quality of the heterologous target complexes produced [17–19, 23, 41]. The MultiBac system furthermore comprises an array of small custom-designed DNA plasmid modules that facilitate the assembly of multigene expression cassettes and their integration into the baculoviral genome (Fig. 13.1). Integration of the multigene expression cassette constructions into the baculoviral genome occurs via two sites (Fig. 13.1). One is a mini-Tn7 attachment site embedded in a LacZcbα gene that is used for blue/white selection and is accessed by the Tn7 transposase which is expressed in the DH10MB cells from a helper plasmid as described previously [43]. Upon integration into this Tn7 site, the LacZα gene is disrupted; white colonies indicate successful transposition. A second entry site is formed by a short imperfect inverted repeat, LoxP, at a location distal from the Tn7 attachment site (Fig. 13.1). It can be accessed by means of the Cre enzyme, a site-specific recombinase that targets the LoxP imperfect inverted repeat (Fig. 13.1). Cre integration occurs by fusing LoxP sites present on the MultiBac genome on the one hand, and on a DNA plasmid module on the other. Successful Tn7 and also Cre integration is imposed by antibiotic selection against the resistance markers encoded by the DNA plasmid modules integrated into the MultiBac genome. The integration sites can be used to integrate genes encoding for one or several multiprotein complexes of choice, but also for additional functionalities that may be required to activate or inactivate the complex (kinases, phosphatases, acetylases, deacetylases, others), support its folding (chaperones) or post-translational processing such as glycosylation [19, 23, 44–46].

The composite MultiBac baculoviral genome which contains all desired heterologous genes is then purified from small bacterial cultures using standard alkaline lysis protocols and applied to small insect cell cultures, typically in six-well plates, with a lipidic or non-lipidic transfection reagent [24, 41]. The resulting live baculovirions are harvested and applied to larger insect cell cultures for heterologous protein production and purification. Production baculovirus is then stored for example at 4 °C in the dark to avoid degeneration of viral titers. A more secure long-term storage method is provided by freezing small aliquots of baculovirus infected insect cells (BIICs) that are then stored in liquid nitrogen [47].

A centerpiece of the MultiBac system is the facilitated assembly of genes into multiexpression cassettes (Figs. 13.1 and 13.2). Originally, this was addressed by creating two different DNA plasmids that contained a so-called Multiplication Module each. This module allowed iterative assembly of single or dual expression cassettes, each fitted with a promoter (p10 or polh) by restriction/ligation utilizing compatible sites that would be destroyed upon ligation [17]. This functional unit-based plasmid configuration later was popularized as ‘BioBrick’ in the context of synthetic biology.

One plasmid (pFBDM) accessed the Tn7 site, the second plasmid (pUCDM) accessed the LoxP site on the MultiBac genome. Both plasmids could be fitted with one to many foreign genes by means of multiplication. The pUCDM plasmid contains a conditional origin of replication derived from the phage R6Kγ. Its survival therefore hinges on the presence of a particular gene (*pir*) in the genome of specific *E. coli* strains. If a pUCDM plasmid is transformed into DH10MB cells which are *pir* negative, it will only survive if it is productively fused to the MultiBac genome by Cre-LoxP reaction [17].

The plasmid module repertoire of the MultiBac system was subsequently expanded by fitting out

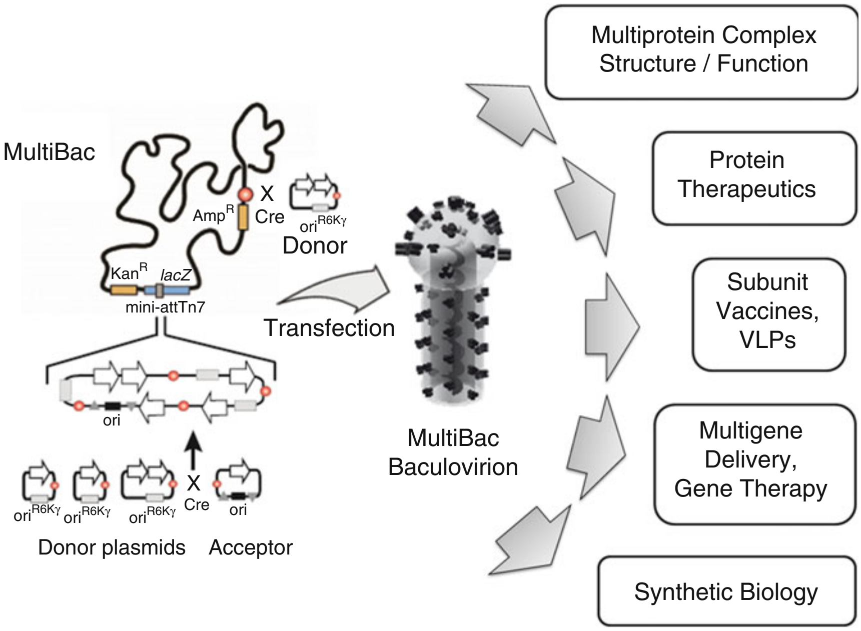


Fig. 13.1 The MultiBac baculovirus/insect cell expression system. The MultiBac system is shown in a schematic view (*left*). MultiBac consist of an engineered baculovirus optimized for protein complex production. This MultiBac baculoviral genome exists as a bacterial artificial chromosome (BAC) in *E. coli* cells. It contains two integration sites for foreign genes, by Tn7 transposition or, alternatively, by site-specific recombination mediated by the Cre enzyme. MultiBac further consists of an array of plasmids called Acceptors and Donors that facilitate multigene assembly. MultiBac baculovirions (*center*) are generated by transfecting composite MultiBac BAC in insect cells. MultiBac is successfully used for a wide variety of applications in basic and applied research and development (*right*). Genes of interest are shown as *arrows filled in white. Circles filled in red* indicate LoxP sites. Origins of replication appear as rectangular shapes. *R6Kγ* phage-derived replicon, *VLP* virus-like particle, *Kan* kanamycin marker, *Amp* Ampicillin marker, *Cre* Cre-LoxP fusion reaction, *LacZα* blue/white selection cassette, *attn7* Tn7 attachment site

the pFBDM plasmid with a LoxP site, giving rise to pFL. A further plasmid was created, pKL, which in contrast to the high-copy number pFL plasmid is propagated at low copy numbers, and thus could accommodate difficult or very large genes that had turned out to be unstable in pFL. A new version of pUCDM was designed with a different resistance marker (spectinomycin), pSPL. pSPL or pUCDM could now be fused either with the MultiBac genome *in vivo* in DH10MB if Cre was expressed, or *in vitro* with pFL or pKL. All plasmids retained the Multiplication Module and therefore could be outfitted with several to many genes in an iterative way. pUCDM and pSPL were now denoted Donor plasmids (D), while pFL and pKL were denoted Acceptors (A) (Fig. 13.2) [18, 41]. One Acceptor could be fused with one or two Donors by Cre-LoxP reaction *in vitro*, and productive fusions were identified by the combination of resistance markers present on the fusion. The fused AD or ADD constructs carrying several to many heterologous genes, could then be inserted into the Tn7 attachment site of the MultiBac genome by transposition in DH10MB cells as before. *In vitro* fusion of AD or ADD plasmids prior to integration into the MultiBac genome via the Tn7 site did not rule out use of the LoxP site present on the viral backbone to additionally incorporate a Donor. Integration into the viral LoxP site simply had to precede the transposition reaction [18].

A *sine qua non* of contemporary structural biology is automation, to increase the throughput of expression experiments by robotic approaches. As a consequence, the multigene assembly technology of the MultiBac system was adapted to

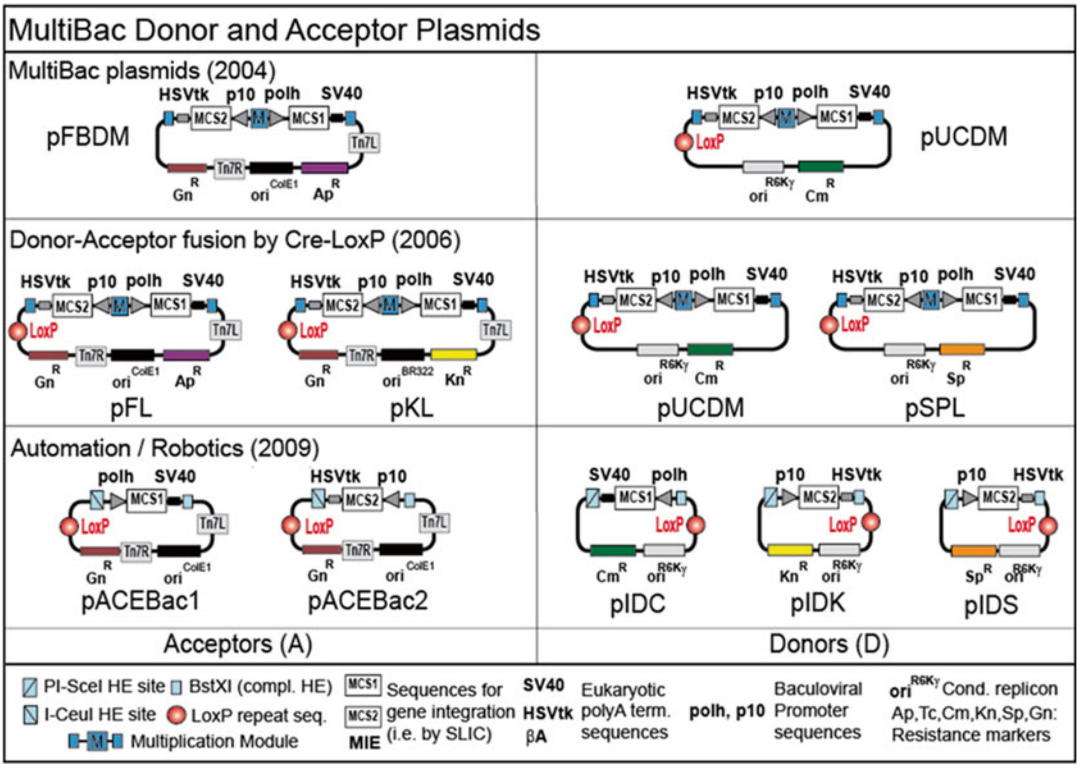


Fig. 13.2 MultiBac tool-kits. A variety of entry plasmids to integrate heterologous genes into the MultiBac baculoviral genome have been created since the introduction of the system in 2004, each with its own merits. Functional modules contained in the plasmids are listed (*bottom*). All plasmids are compatible with each other and can be used in various combinations to generate recombinant MultiBac baculoviral genomes for multiprotein expression and/or multigene delivery. Expression cassettes are in 'BioBrick' format and enable iterative multiplication

automation by using a liquid handling robot [24, 25]. For this, the tandem recombineering technology (TR) originally developed for prokaryotic complex expression [38] was adapted to the MultiBac system [48]. TR combines sequence-and-ligation independent gene insertion (SLIC) with Cre-LoxP fusion to generate multigene expression constructs in high-throughput in a robotic setup. Adaptation of the MultiBac plasmids to TR required subtle adjustments of the plasmids resulting in Acceptor plasmids pACEBac1 and pACEBac2 as well as Donor plasmids pIDK, pIDS and pIDC that are robotics-compatible [22, 24, 40]. Concomitantly, with the objective to simplify the monitoring of protein production by measuring fluorescence levels, a new baculovirus genome, EMBacY, was created expressing a yellow fluorescent protein (YFP) from its backbone, with fluorescence intensity increasing in parallel with the quantity of the heterologous protein complex expressed at the same time.

All Acceptor and Donor plasmids developed for the MultiBac system over time are compatible with each other, and also with the MultiBac and EMBacY genomes (and other MultiBac genome derivatives) that were and are being developed (see also Sects. 13.2.3 and 13.2.4).

13.2.2 A MultiBac User Access Platform

The MultiBac system rapidly developed into a sought after tool following its introduction and original publication, compellingly underscoring

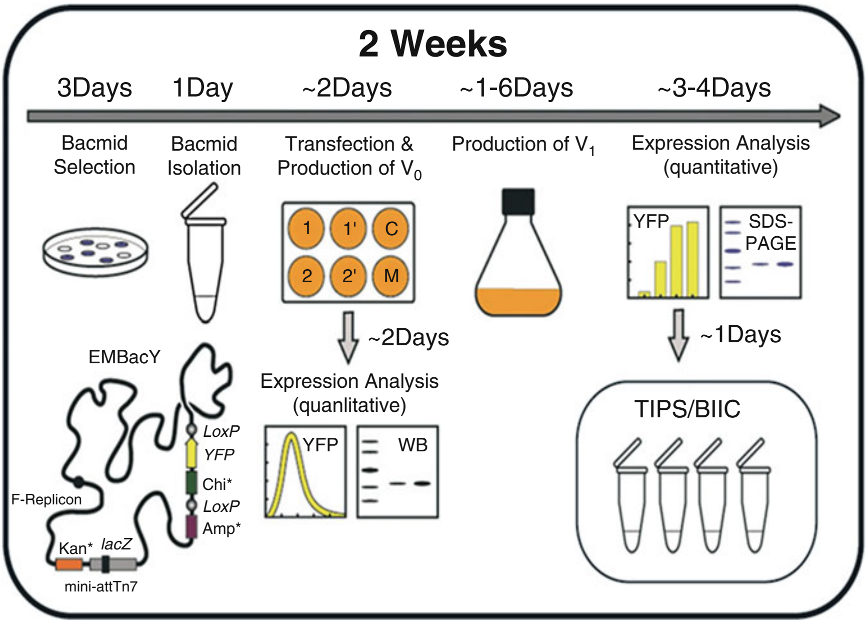


Fig. 13.3 MultiBac expression platform at the EMBL Grenoble. The standard operating protocol (SOP) implemented is illustrated, to express proteins and their complexes by MultiBac. The entire process takes 2 weeks from generation of the composite MultiBac BAC (bacmid) to quantitative expression analysis. A MultiBac baculovirus variant called EMBacY is used in the platform, producing yellow fluorescent protein (YFP) to track virus performance and heterologous protein production. In addition, protein production is monitored by Western blot (WB) analysis or by gel electrophoresis (SDS-PAGE). Production virus is stored long-term in frozen aliquots of baculovirus in insect cells (BIICs)

the current need for an accessible expression technology for eukaryotic multiprotein complexes. The number of laboratories using MultiBac approaches a thousand at the time this review is written. Research groups in academia as well as the biotech and pharma industries are implementing MultiBac to produce their specimens of interest. Moreover, biotech spin-offs were founded based on MultiBac developments, including a preclinical vaccine development company, Redvax GmbH, which in 2015 was acquired by the global pharma enterprise, Pfizer.

The high demand for accessing this technology resulted in the establishment of a dedicated research and training platform at the European Molecular Biology Laboratory (EMBL) in the Eukaryotic Expression Facility (EEF) established at the EMBL Outstation in Grenoble (Fig. 13.3). The EEF has been supported since 2008 by the European Commission (EC) through infrastructure grants (P-CUBE, BioStruct-X, INSTRUCT) and by the national (French) research agency (ANR) through the Investissement d'Avenir program. More than a hundred projects per year, by local national and transnational users, including academic research projects as well as industrial contracts have been processed in the facility, covering a wide range of applications in basic and applied research and development. Academic project access is based on the sole criteria of excellence, determined by an independent panel reviewing research proposals. The facility has implemented a SOP-based procedure for routinely and rapidly moving projects through the MultiBac platform pipeline [23, 24].

The entire process from preparing the multigene expression construct to (small-scale) purification of the specimens of interest requires around 2 weeks according to this protocol. A large number of constructs can be processed in parallel. Virus amplification as well as heterologous protein production is monitored by measuring the signal of yellow fluorescent protein (YFP) in small (one million cells) aliquots withdrawn from the cell cultures at defined intervals. The YFP signal reaching a plateau indicates maximal production of the desired complex specimen and the expression culture is harvested at this point for further processing and purification. Particular care is taken to assure highest quality virus production during virus amplification—early (budded) virus is harvested throughout the amplification process to avoid accumulation of defective viral particles that would compromise virus performance. Initial virus is stored at 4 °C, while production virus is stored as BIICs in liquid nitrogen [21–24] (Fig. 13.3).

13.2.3 OmniBac: Universal Multigene Transfer Vectors

Two approaches for foreign gene integration into the baculoviral genome dominate the field. One approach depends on the baculoviral genome present as a BAC in *E. coli* cells, and relies on gene integration by transposition catalyzed by the Tn7 transposase which is constitutively expressed from a dedicated (helper) plasmid in the cells harboring the BAC. The foreign genes are provided by transforming a transfer plasmid into these *E. coli* cells. Selection for recombinant BACs occurs based on the resistance marker that is present on the transfer plasmid and also integrated, as well as blue/white selection; productive transposase mediated integration destroys a LacZα gene. The MultiBac system retains this strategy, extended by the option to provide in addition to the Tn7 transposase also the site-specific Cre recombinase transiently from a second dedicated helper plasmid, to fuse an additional Donor construct into a LoxP site present on the MultiBac baculoviral genome [17, 40, 41].

The alternative, original approach is based on homologous recombination, mediated by DNA regions in the transfer plasmid that are also present in the baculoviral genome, flanking the polh locus that had been inactivated by destruction of the *polh* gene. These regions of homology correspond to the open reading frames Orf1620 and lef2/603. By using this method, insertion occurs by co-transfecting the transfer plasmid and the purified baculoviral genome into insect cells. The baculoviral genome is typically linearized in the region between Orf1629 and lef2/603, to suppress formation of virus devoid of the heterologous DNA of interest. Fusion of the plasmid DNA with the baculoviral DNA to create a replicating genome is then achieved via the homologous recombination system of the insect cells. The genome is thus re-circularized, concomitantly inserting the gene of interest. Live virions are then produced that express the desired protein(s).

Both methods have advantages and disadvantages. The Tn7/BAC based integration approach is the method of choice in many laboratories, mainly due to its simplicity. Since the baculoviral genome exists as a BAC in *E. coli*, it can be, in theory, propagated indefinitely and used for many experiments after obtaining the initial aliquot. Moreover, it can also be manipulated by gene editing technologies in its *E. coli* host cell. In contrast, the linearized baculoviral DNA for homologous recombination has to be obtained from the supplier for every experiment *de novo* and cannot be propagated at will. Furthermore, the homologous recombination method is arguably somewhat more involved and may require specialist knowledge. On the other hand, this approach is more amenable to automation as compared to the BAC-based method which is characterized by many steps some of which (such as blue/white screening) cannot be readily scripted into robotics routines. A further disadvantage of the BAC-based system may be found in the relative instability that was described for BAC-derived baculoviruses in insect cells, presumably originating from the presence of extended bacterial DNA elements (selection marker, F-replicon, LacZα gene) [4, 49], limiting its use in human applications mainly to preclinical studies [4]. In fact, baculoviruses used in commercial production to date are still being

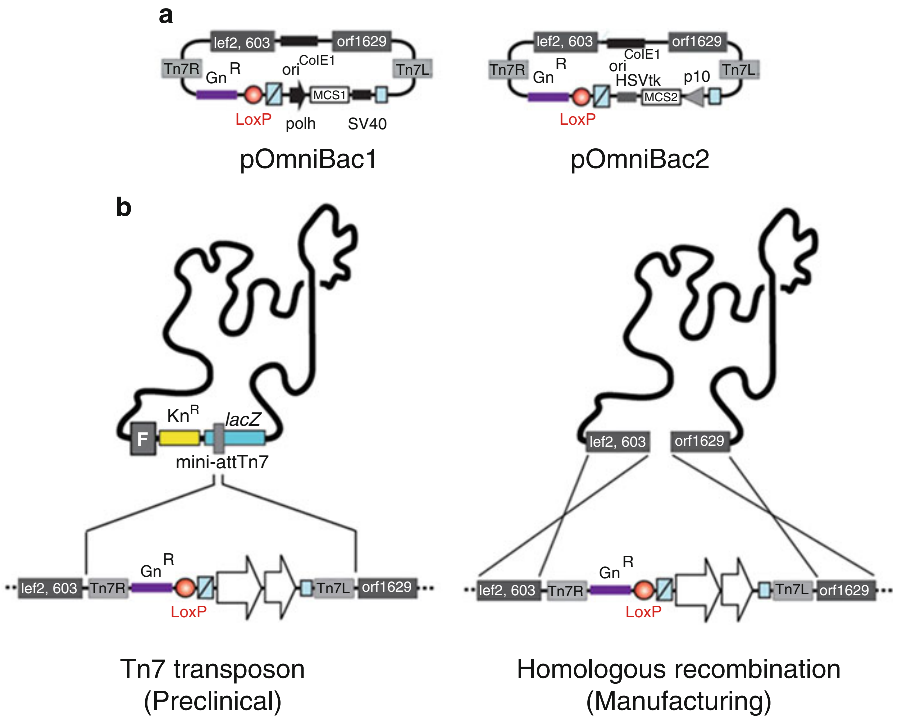


Fig. 13.4 OmniBac – Universal transfer plasmids for every BEVS. (**a**) Acceptor plasmids pOmniBac1 and pOmniBac2 are shown schematically, functional modules are same as listed above (Fig. 13.2, bottom). These Acceptors can be combined with the Donor plasmids by Cre-LoxP recombination. (**b**) OmniBac plasmids comprise elements for homologous recombination as well as Tn7-based transposition. Multigene constructions based on OmniBac plasmids can therefore access all available baculovirus genomes. Thus, with the same plasmids, composite baculovirus for preclinical studies as well as manufacturing can be produced efficiently

made almost exclusively by applying the classical homologous recombination technique [4].

Available BEVS all relied on either one method or the other, which were mutually exclusive, and the choice of the transfer plasmid decided which virus would be used for protein production. This situation is unsatisfactory given that both systems (BAC/Tn7-based and classical) provide unique opportunities. Moreover, numerous baculoviruses with customized functionalities have been created for both applications, each with its own merit. We therefore designed and created the "OmniBac" transfer plasmids which combine the DNA elements required for Tn7 transposition in the BAC-based system and also the homology regions for baculovirus generation following the classical method, and thereby are universally applicable (Fig. 13.4). These OmniBac transfer plasmids function as Acceptors in the MultiBac system (Fig. 13.4). They can be combined with a variety of Donors to yield multigene Acceptor-Donor fusions that can then be funneled into any baculovirus of choice [50].

13.2.4 The ComplexLink Polyprotein Expression Technology

A challenge frequently encountered when overexpressing protein complexes relates to imbalanced expression levels of the individual protein subunits that are to assemble into the biological target superstructure. If a particular subunit is badly made in a co-expression experiment, it will limit the overall yield of the fully assembled protein complex dramatically. This challenge can

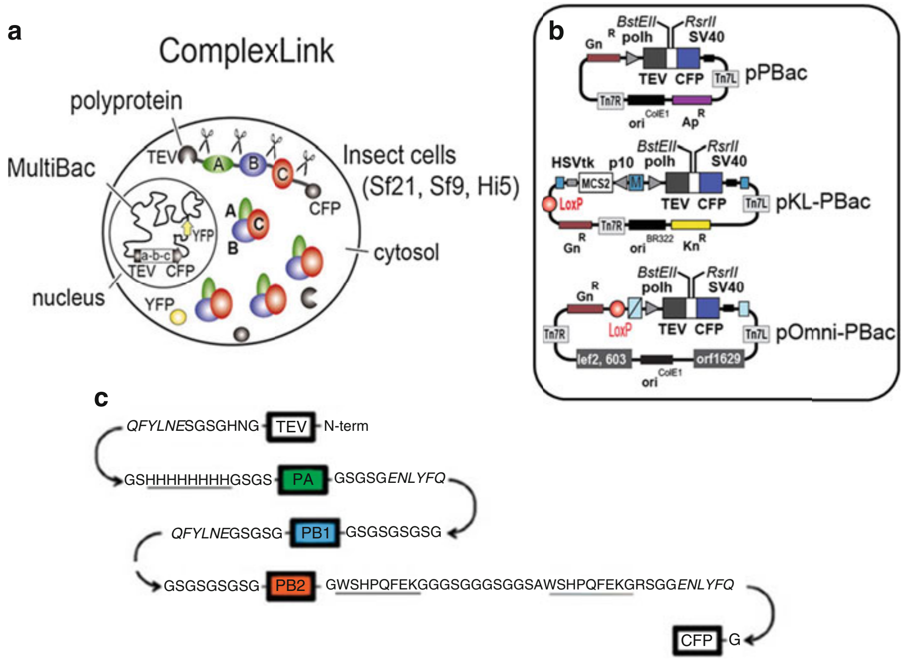


Fig. 13.5 ComplexLink technology. (**a**) The ComplexLink technology was created to produce multiprotein complexes from self-processing polyprotein constructs as shown here schematically. The polyprotein contains a TEV protease and a fluorescent protein, at the N- and C-termini, respectively. Polyproteins are processed into the individual protein entities by the highly specific TEV protease. (**b**) Polyprotein expression plasmids pPBac, pKL-PBac and pOmni-PBac are shown. DNA modules are marked as above (Fig. 13.2, *bottom*). **c** Schematic representation of a self-processing polyprotein encoding for influenza polymerase, before TEV-mediated cleavage. TEV stands for tobacco etch virus NIa protease, PA, PB1 and PB2 are subunits of the trimeric influenza enzyme complex, CFP stands for cyan fluorescent protein. BstEII and RsrII are asymmetric restriction enzymes that can be used to access polyprotein expression plasmids for restriction/ligation-based heterologous gene integration

be addressed within limits by co-infection approaches using several viruses, or the choice of promoters. In contrast to very late promoters such as polh and p10, earlier promoters are expressed at lower levels. However, it may be often impractical to resort to co-infection or to tuning protein expression levels by promoter choice, also due to the fact that the timing of the production of protein subunits will then be altered as well, with often unpredictable consequences. A solution to such problems can derive from observing the strategies certain viruses utilize to realize their proteome. Coronavirus, the agent that causes severe acute respiratory syndrome (SARS), produces its complete proteome from two long open reading frames (ORFs) that give rise to polyproteins in which the individual protein specimens are covalently linked. A highly specific protease, also encoded by the ORF, then liberates the individual proteins by cleaving them apart.

The ComplexLink technology implements this strategy for recombinant polyprotein production from polygenes [22, 40, 51] (Fig. 13.5). In ComplexLink, genes encoding for a protein complex of choice are conjoined to yield a single ORF. This ORF gives rise to a polyprotein in which the individual proteins are linked by short amino acid sequences representing a cleavage site for the NIa protease from tobacco etch virus (TEV) which is also encoded by the ORF and is the first protein produced. In addition, a cyan fluorescent marker protein (CFP) is present at the C-terminal end of the polyprotein construction. Upon translation, TEV protease liberates itself, and all other proteins including CFP by cleaving

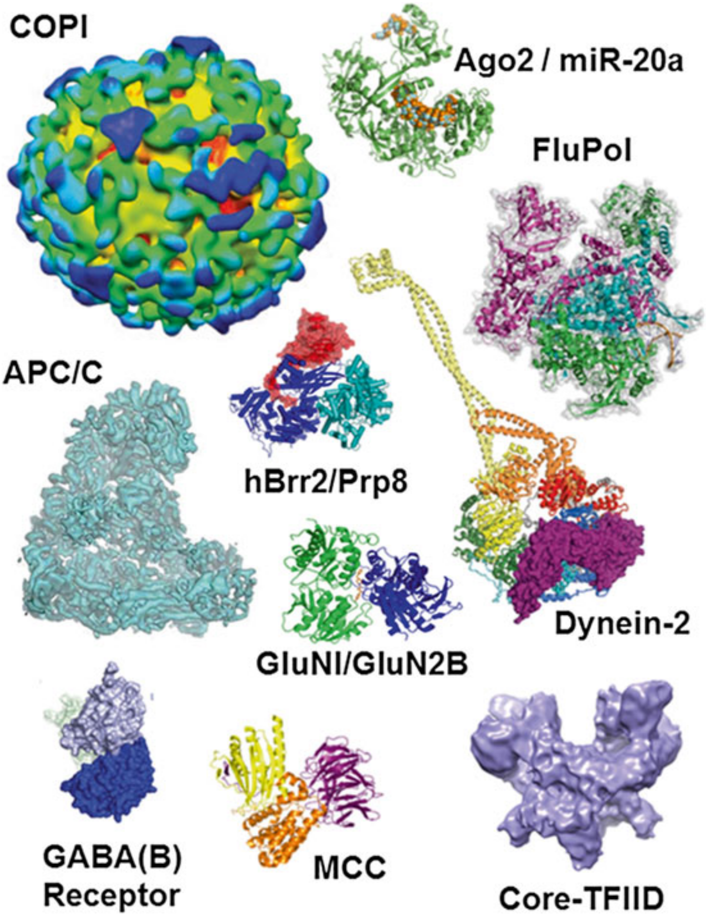


Fig. 13.6 MultiBac complex structure gallery. A selection of recent high impact structures of MultiBac-produced biological specimens are shown. Examples include cryo-EM architectures of COPI-coated Vesicles (EMD-2084 to EMD-2088), the complete human APC/C complex with coactivator and substrate (EMD-2651 to EMD-2654) and the human core-TFIID complex (EMD-2229 to EMD-2231). Notable structures that were determined by X-ray crystallography (PDB identifiers are provided in brackets) include influenza polymerases A and B bound to the viral RNA promoter (PDB identifier 4WSA, 4WRT), human Argonaute Ago2 in complex with miR-20a RNA (4F3T), the spliceosomal complex Brr2^{HR}/Prp8^{Jab1} (4KIT), human GABA(B) receptor (PDB 4MQE), the dynein-2 motor bound to ADP (4RH7), the mitotic checkpoint complex MCC (4AEZ) and the GluN1/GluN2B *N*-methyl-D-aspartate receptor (3QEL). Molecular illustrations were prepared with PyMOL (www.pymol.org) and Adobe Photoshop Version CS6

the TEV-specific proteolytic sites. The ComplexLink plasmids, pPBac, pKL-PBac and pOmni-PBac function as Acceptors in the MultiBac system and can be fused to Donors which may contain further genes encoding for polyproteins. The ComplexLink technology proved to be highly successful in producing difficult-to-express protein complexes in high quality and quantity, including the physiological core complex of human general transcription factor TFIID [52–54]. A notable exploit is influenza polymerase, an important drug target to combat the flu, which has remained inaccessible for 40 years since its discovery. Influenza polymerase has been produced, for the first time, successfully using ComplexLink in conjunction with MultiBac, enabling elucidation of its structure and mechanism by X-ray crystallography at near-atomic resolution [55, 56] (Figs. 13.5 and 13.6).

13.3 Applications

The MultiBac system in its original form was introduced in 2004, and has become the method of choice for a wide range of applications. Primarily developed for accelerating structural biology of multiprotein complexes, it has since then been modified and improved to benefit also other fields, in basic and applied research. We have followed these developments with interest and have occasionally highlighted them in invited review articles and commentaries [40, 53, 57, 58]. In the following, we intend to summarize these developments without being exhaustive, focusing on recent exploits by researchers who adopted the system we had developed, to catalyze their research.

13.3.1 Accelerating Complex Structural Biology

The first users which implemented MultiBac were structural biologists interested in elucidating the architecture and mechanics of important multiprotein machines in cell biology at the molecular, near-atomic level. This is also the field where an impressive flurry of highest impact MultiBac-enabled contributions was achieved to date. We had highlighted some of these exciting structures, including important drug targets such as the LKB1-STRAD-MO25 complex [59], and the first structure of a nucleosome-bound chromatin remodeler, Isw1 [60] in a contribution just over 2 years ago in *Trends in Biochemical Sciences* [40]. In the short time since then, spanning a mere 2 years, a large number of new structural studies were carried out using material produced with MultiBac. A selection of these exploits is presented in Fig. 13.6.

Landmark achievements are the recent crystal structures of influenza polymerase [55, 56]. This success was catalyzed by applying the ComplexLink and MultiBac technologies in combination, to produce this trimeric protein complex which had remained elusive for decades. The structures describe fluA and fluB variants of the polymerase bound to its RNA ligand, and provide important structural insights into cap-snatching and RNA synthesis by this enzyme complex, opening up new avenues for pharmaceutical development to combat flu. Further crystallographic exploits include the structures of human cytoplasmic dynein-2 primed for its power stroke [61], the human argonaute-2/RNA complex [62], the structure of the spliceosomal protein Prp8 bound to an RNA helicase, Brr2 [63, 64], and structures of PI4KIIIβ kinase complexes [65] among numerous others. Highlights achieved by using MultiBac produced material for electron microscopic studies include the structures of COP1-coated vesicles, revealing alternate coatomer conformations and interactions [66, 67], the architecture of the physiological core of human general transcription factor TFIID [52, 58], or the elucidation of the molecular mechanisms of the anaphase promoting complex APC/C at sub-nanometer resolution [68]. Recently, the entire human Mediator transcription factor holo-complex has been successfully assembled by using MultiBac, and functionally characterized [69]. MultiBac reagents have been incorporated into pipelines for producing membrane proteins and their complexes [70]. We anticipate that many more exciting structures of important protein assemblies will be determined in the future, by using the MultiBac system as a production tool.

13.3.2 MultiBac in Pharma and Biotech

The baculovirus/insect cell system has had a major impact on the production of high-value protein targets, for pharmacological characterization, structure-based drug design, diagnostics, biosensor engineering and high-throughput proteomics [2, 4, 5, 71]. Notably human proteins, virus-like particles (VLPs) and vaccines have been successfully expressed by using BEVS [2, 4, 5]. Glycoproteins are sought-after biologics in the pharma and biotech sector, and insect cells have proven to be well suited for the expression of biologically active and immunogenic specimens. The MultiBac system has been engineered to enable high-quality production of glycoproteins and their complexes. The original MultiBac baculoviral genome was already lacking the *v-cath* and *chiA* genes, which are encoding for cathepsin-like protease and chitinase, respectively. Both v-cath and chiA have been shown to be detrimental to glycoprotein production [72–74]. The glycosylation pattern of secreted proteins in insect cells differs from mammalian patterns which involve more complex N-glycans [75, 76]. These differences can have adverse effects on human glycoproteins produced in insect cells. To overcome this impediment, a new MultiBac-derived baculovirus, SweetBac, was constructed, which includes glycosyltransferases in the backbone, resulting in mammalianized glycosylation patterns of SweetBac-produced glycoproteins [23, 44, 45]. More recently, improved MultiBac variants were introduced to minimize fucosylation in insect cell derived glycoproteins to reduce binding to antibodies from the sera of patients with allergies [46].

The BEVS has demonstrated its aptitude to produce complex multicomponent assemblies such as virus-like particles (VLPs) [4, 9–11, 77]. VLPs are promising candidates for vaccination. VLPs resemble natural virus shells, but are lacking genetic material and therefore are safe and not infectious. VLPs can be proteinaceous, such as for example papilloma VLPs used to prevent cervical cancer. More recently, enveloped viruses have been successfully produced using BEVS, including influenza and chikungunya vaccine candidates [4, 77–80]. The MultiBac system was already successfully used to produce a number of VLPs including an array of papilloma serotypes [40]. Complex VLPs representing highly pathological virus strains were produced safely [81]. In particular the availability of OmniBac plasmids, which are part of the MultiBac vector suite, may provide unique opportunities for VLP vaccine development, as they are equally useful for exploratory preclinical studies in high-throughput as well as pharmacological manufacturing, by choosing the most appropriate viral backbone for large scale expression (Fig. 13.4).

Baculoviruses not only infect insect cells, but can also transduce mammalian cells efficiently [15, 82–84]. By choosing mammalian-active promoters instead of polh, p10 or other baculoviral promoters, proteins of interest can be produced from a baculovirus that has entered a mammalian host cell. Baculoviruses do not replicate in mammalian cells, therefore, the current consensus is that this BacMam approach can be performed safely in laboratories. A particular benefit of BacMam is that large DNA insertions including multicomponent signaling cascades or entire metabolic pathways can be transduced into mammalian cells by the baculovirus which can tolerate very large gene insertions, and can be amplified and produced in large amount in a straightforward manner in insect cell cultures. Baculovirus is thus emerging as a highly promising gene delivery tool into mammalian cells, for a multitude of applications [14, 15]. Already, multigene MultiBac constructions were successfully used to produce recombinant adeno-associated viruses (AAVs) for gene therapy [12, 13, 40].

13.3.3 Synthetic Biology: Rewiring the Genome

The AcMNPV genome has a size of around 130 kilobases and contains numerous functionalities, which are essential in the natural life cycle of the virus, but dispensable in laboratory culture. Moreover, in the laboratory, it has been recognized that the genome has a tendency to undergo multiple deletions in its genome during amplifications, notably in regions containing foreign gene cargo that has been inserted for overproduction [4, 50]. This can have severely detrimental consequences for heterologous target protein production, especially for fermenter-scale manufacturing of biologics which require large foreign DNA insertions in the viral genome and several rounds of amplification, until sufficient volumes of production virus are obtained to charge the fermenter. We have shown that some of these limitations can be overcome at least on laboratory-scale by stringently adhering to virus generation protocols that avoid or limit the occurrence of wide-spread deletions [18]. Moreover, it appears that there is considerable scope for improving virus performance by eliminating mutational or deletion hot-spots in the viral genome, and by removing DNA regions which are not required in cell culture. The baculoviral genome, in particular when present as a BAC, lends itself excellently to genome manipulation and editing techniques. We and others have exploited this avenue to remove unnecessary or undesired functionalities from the virus, such as the *polh*, *p10*, *v-cath* and *chiA* genes among (many) others.

The advent of efficient synthetic biology techniques now holds the promise to reverse this somewhat cumbersome top-down approach, and to rationally redesign and rewire the baculoviral genome bottom-up by applying DNA synthesis and assembly methods that have become available more recently. Interestingly, comparison of baculovirus sequences in genome databases indicate that most genes and DNA elements thought to be required for survival of the virus in cell culture are confined to roughly one half of the circular viral genome, while the other half contains DNAs that can be probably largely disposed of in the laboratory [50]. This approach may hold challenges given the complex interplay of baculoviral proteins and their relative expression levels during the different phases of the viral life cycle [4]. Notwithstanding, synthetic approaches, probably best applied in combination with sequential deletions, are exciting and potentially highly rewarding avenues for developing new and minimal baculoviral genomes that can be customized for optimal properties in the research laboratory and also in industrial manufacturing.

13.4 Outlook

Since the pioneering first reports more than three decades ago, the baculovirus/insect cell expression system has developed into a mainstream production platform, accelerating a wide range of research projects in academic and industrial laboratories. In the post-genomic era, multiprotein complexes have entered center stage as essential catalysts of cellular activity, and notably the MultiBac BEVS has contributed substantially to make hitherto inaccessible protein complexes available, to unlock their structure and mechanism in molecular detail. Moreover, BEVS has emerged as a remarkably useful tool in the biotech and pharma sector, for the production of complex biologics in disease prevention and therapy. The development of this versatile expression tool is continuing unabated as it is set to benefit markedly from powerful new synthetic biology techniques that are becoming readily available. Fueled by these innovations, BEVS is excellently positioned to play a key and increasing role in the life sciences, in basic and applied research in the future. Exciting times are ahead of us.

Acknowledgments We thank all members of the Berger laboratory for their support, and in particular Lakshmi S. Vijayachandran (Amrita Center for Nanosciences & Molecular Medicine, India), Yan Nie (Shanghai Institute of Technology, China), Simon Trowitzsch (Goethe University, Frankfurt) and Christoph Bieniossek (Hofmann-La Roche, Basel) for their contribution and helpful discussions. This work was funded by the European Commission (EC) Framework Programme (FP) 7 ComplexINC project (grant number 279039), to DF and IB. The MultiBac user access platform at the EMBL Grenoble received funding from the EC through the FP6 and FP7 projects SPINE2C, P-CUBE, BioStruct-X and INSTRUCT (grant numbers 031220, 227764, 283570 and 211252), and from the French Agence Nationale de Recherhe (ANR) through the Investissement d'Avenir funding scheme. DBTGR was a fellow of the EMBL interdisciplinary post-doctoral opportunities program (EIPOD).

References

1. Smith GE, Summers MD, Fraser MJ (1983) Production of human beta interferon in insect cells infected with a baculovirus expression vector. Mol Cell Biol 3:2156–2165
2. Summers MD (2006) Milestones leading to the genetic engineering of baculoviruses as expression vector systems and viral pesticides. Adv Virus Res 68:3–73
3. Pennock GD, Shoemaker C, Miller LK (1984) Strong and regulated expression of Escherichia coli beta-galactosidase in insect cells with a baculoviral vector. Mol Cell Biol 4:399–406
4. van Oers MM, Pijlman GP, Vlak JM (2015) Thirty years of baculovirus-insect cell protein expression: from dark horse to mainstream technology. J Gen Virol 96:6–23
5. Contreras-Gomez A, Sanchez-Miron F, Garcia-Camacho F, Molina-Grima E, Chisti Y (2014) Protein production using the baculovirus-inseect cell expression system. Biotechnol Prog 30:1–18
6. Jarvis DL (2009) Baculovirus-insect cell expression systems. Methods Enzymol 463:191–222
7. Possee RD, King LA (2007) Baculovirus transfer vectors. Methods Mol Biol 388:55–76
8. Perez de Diego AC et al (2011) Characterisation of protection afforded by a bivaleent virus-like partcile vaccine against bluetongue virus serotypes 1 and 4 in sheep. PLoS One 6:e26666
9. Vicente T, Roldao A, Peixeto C, Carrondo MJT, Alves PM (2011) Large-sclae production and purification of VLP-based vaccines. J Invertebr Pathol 107(Suppl):S42–S48

10. Roy P, Noad R (2009) Bluetongue vaccines. Vaccine 27(Suppl):D86–D89
11. Roy P, Noad R (2009) Virus-like partciles as a vaccine delivery system: myths and facts. Adv Exp Med Biol 655:145–158
12. Mietzsch M et al (2014) OneBac: platform for scalable and high-titer production of adeno-associated virus serotype 1–12 vectors for gene therapy. Hum Gene Ther 25(3):212–222
13. Marsic D et al (2014) Vector design Tour de Force: integrating combinatorial and rational approaches to derive novel adeno-associated virus variants. Mol Ther 22(11):1900–1909
14. Kotin RM (2011) Large-scale recombinant adeno-associated virus production. Hum Mol Genet 20(R1):R2–R6
15. Airenne KJ et al (2013) Baculovirus: an insect-derived vector for diverse gene transfer applications. Mol Ther 21(4):739–749
16. Paul A, Hasan A, Rodes L, Sangaralingam M, Prakash S (2014) Bioengineered baculoviruses as new class of therapeutics using micro and nanotechnologies: principles, prospects and challenges. Adv Drug Deliv Rev 71:115–130
17. Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. Nat Biotechnol 22(12):1583–1587
18. Fitzgerald DJ et al (2006) Protein complex expression by using multigene baculoviral vectors. Nat Methods 3(12):1021–1032
19. Fitzgerald DJ et al (2007) Multiprotein expression strategy for structural biology of eukaryotic complexes. Structure 15(3):275–279
20. Bieniossek C, Berger I (2009) Towards eukaryotic structural complexomics. J Struct Funct Genomics 10(1):37–46
21. Trowitzsch S, Bieniossek C, Nie Y, Garzoni F, Berger I (2010) New baculovirus expression tools for recombinant protein complex production. J Struct Biol 172(1):45–54
22. Vijayachandran LS et al (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J Struct Biol 175(2):198–208
23. Trowitzsch S, Palmberger D, Fitzgerald D, Takagi Y, Berger I (2012) MultiBac complexomics. Expert Rev Proteomics 9(4):363–373
24. Berger I et al (2013) The MultiBac protein complex production platform at the EMBL. J Vis Exp 11(77):e50159
25. Berger I, Chaillet M, Garzoni F, Yau-Rose S, Zoro B (2013) High-throughput screening of multiple protein complexes. Am Lab 25(8):32–35
26. Nie Y et al (2009) Getting a grip on complexes. Curr Genomics 10(8):558–572
27. Robinson CV, Sali A, Baumeister W (2007) The molecular sociology of the cell. Nature 450(7172):973–982
28. Ramakrishnan V (2014) The ribosome emerges from a black box. Cell 159(5):979–984
29. Fernández-Tornero C et al (2013) Crystal structure of the 14-subunit RNA polymerase I. Nature 502(7473):644–649
30. Engel C, Sainsbury S, Cheung AC, Kostrewa D, Cramer P (2013) RNA polymerase I structure and transcription regulation. Nature 502(7473):650–655
31. Cramer P et al (2008) Structure of eukaryotic RNA polymerases. Annu Rev Biophys 37:337–352
32. Völkel P, Le Faou P, Angrand PO (2010) Interaction proteomics: characterization of protein complexes using tandem affinity purification-mass spectrometry. Biochem Soc Trans 38(4):883–887
33. Li Y (2010) Commonly used tag combinations for tandem affinity purification. Biotechnol Appl Biochem 55(2):73–83
34. Janin J, Séraphin B (2003) Genome-wide studies of protein-protein interaction. Curr Opin Struct Biol 13(3):383–388
35. Almo SC et al (2013) Protein production from the structural genomics perspective: achievements and future needs. Curr Opin Struct Biol 23(3):335–344
36. Haffke M et al (2015) Characterization and production of protein complexes by co-expression in Escherichia coli. Methods Mol Biol 1261:63–89
37. Vincentelli R, Romier C (2013) Expression in Escherichia coli: becoming faster and more complex. Curr Opin Struct Biol 23(3):326–334
38. Bieniossek C et al (2009) Automated unrestricted multigene recombineering for multiprotein complex production. Nat Methods 6(6):447–450
39. Abdulrahman W et al (2015) The production of multiprotein complexes in insect cells using the baculovirus expression system. Methods Mol Biol 1261:91–114
40. Bieniossek C, Imasaki T, Takagi Y, Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 37(2):49–57
41. Bieniossek C, Richmond TJ, Berger I (2008) MultiBac: multigene baculovirus-based eukaryotic protein complex production. Curr Protoc Protein Sci 51:5.20.1 - 5.20.26
42. Tsutsui H, Matsubara K (1981) Replication control and switch-off function as observed with a mini-F factor plasmid. J Bacteriol 147(2):509–516
43. Luckow VA, Lee SC, Barry GF, Olins PO (1993) Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in *Escherichia coli*. J Virol 67:4566–4579
44. Palmberger D, Wilson IB, Berger I, Grabherr R, Rendic D (2012) SweetBac: a new approach for the production of mammalianised glycoproteins in insect cells. PLoS One 7(4):e34226
45. Palmberger D, Klausberger M, Berger I, Grabherr R (2013) MultiBac turns sweet. Bioengineered 4(2):78–83

46. Palmberger D et al (2014) Minimizing fucosylation in insect cell-derived glycoproteins reduces binding to IgE antibodies from the sera of patients with allergy. Biotechnol J 9(SI):1206–1214
47. Wasilko DJ et al (2009) The titerless infected-cells preservation and scale-up (TIPS) method for large-scale production of NO-sensitive human soluble guanylate cyclase (sGC) from insect cells infected with recombinant baculovirus. Protein Expr Purif 65(2):122–132
48. Haffke M, Viola C, Nie Y, Berger I (2013) Tandem recombineering by SLIC cloning and Cre-LoxP fusion to generate multigene expression constructs for protein complex research. Methods Mol Biol 1073:131–140
49. Pijlman GP, van Schijndel JE, Vlak JM (2003) Spontaneous excision of BAC vector sequences from bacmid-derived baculovirus expression vectors upon passage in insect cells. J Gen Virol 84:2669–2678
50. Vijachandran LS et al (2013) Gene gymnastics: synthetic biology for baculovirus expression vector system engineering. Bioengineered 4(5):279–287
51. Nie Y, Bellon-Echeverria I, Trowitzsch S, Bieniossek C, Berger I (2014) Multiprotein complex production in insect cells by using polyproteins. Methods Mol Biol 1091:131–141
52. Bieniossek C et al (2013) The architecture of human general transcription factor TFIID core complex. Nature 493(7434):699–702
53. Barford D, Takagi Y, Schultz P, Berger I (2013) Baculovirus expression: tackling the complexity challenge. Curr Opin Struct Biol 23(3):357–364
54. Trowitzsch S et al (2015) Cytoplasmic TAF2-TAF8-TAF10 complex provides evidence for nuclear holo-TFIID assembly from preformed submodules. Nat Commun 6:6011
55. Reich S et al (2014) Structural insight into cap-snatching and RNA synthesis by influenza polymerase. Nature 516(7531):361–366
56. Pflug A, Gulligay D, Reich S, Cusack S (2014) Structure of influenza A polymerase bound to the viral RNA promoter. Nature 516(7531):355–360
57. Berger I, Mary LM (2013) Protein production for structural biology: new solutions to new challenges. Curr Opin Struct Biol 23(3):317–318
58. Kandiah E, Trowitzsch S, Gupta K, Haffke M, Berger I (2014) More pieces to the puzzle: recent structural insights into class II transcription initiation. Curr Opin Struct Biol 24:91–97
59. Zeqiraj E, Filippi BM, Deak M, Alessi DR, van Aalten DM (2009) Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation. Science 326(5960):1707–1711
60. Yamada K et al (2011) Structure and mechanism of the chromatin remodelling factor ISW1a. Nature 472(7344):448–453
61. Schmidt H, Zalyte R, Urnavicius L, Carter AP (2015) Structure of human cytoplasmic dynein-2 primed for its power stroke. Nature 518(7539):435–438
62. Elkayam E et al (2012) The structure of human argonaute-2 in complex with miR-20a. Cell 150(1):100–110
63. Santos KF et al (2012) Structural basis for functional cooperation between tandem helicase cassettes in Brr2-mediated remodeling of the spliceosome. Proc Natl Acad Sci U S A 109(43):17418–17423
64. Mozaffari-Jovin S et al (2013) Inhibition of RNA helicase Brr2 by the C-terminal tail of the spliceosomal protein Prp8. Science 341(6141):80–84
65. Burke JE et al (2014) Structures of PI4KIIIβ complexes show simultaneous recruitment of Rab11 and its effectors. Science 344(6187):1035–1038
66. Sahlmuller MC et al (2011) Recombinant heptameric coatomer complexes: novel tools to study isoform-specific functions. Traffic 12(6):682–692
67. Faini M et al (2012) The structures of COPI-coated vesicles reveal alternate coatomer conformations and interactions. Science 336(6087):1451–1454
68. Chang L, Zhang Z, Yang J, McLaughlin SH, Barford D (2014) Molecular architecture and mechanism of the anaphase-promoting complex. Nature 513(7518):388–393
69. Cevher MA et al (2014) Reconstitution of active human core mediator complex reveals a critical role of the MED14 subunit. Nat Struct Mol Biol 21(12):1028–1034
70. Goehring A et al (2014) Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. Nat Protoc 9(11):2574–2585
71. Drugmand JC, Schneider YJ, Agathos SN (2012) Insect cells as factories for biomanufacturing. Biotechnol Adv 30(5):1140–1157
72. Hom LG, Ohkawa T, Trudeau D, Volkman LE (2002) *Autographa californica* M nucleopolyhedrovirus ProV-CATH is activated during infected cell death. Virology 131:561–565
73. Hitchman RB et al (2010) Improved expression of secreted and membrane-targeted proteins in insect cells. Biotechnol Appl Biochem 56:85–93
74. Kaba SA, Salceda AM, Wafula PO, Vlak JM, van Oers MM (2004) Development of a chitinase and v-cathepsin negative bacmid of improved integrity of secreted recombinant proteins. J Virol Methods 122:113–118
75. Harrison RL, Jarvis DL (2006) Protein N-glycosylation in the baculovirus-insect cell system and engineering of insect cells to produce “mammalianized” recombinant glycoproteins. Adv Virus Res 68:159–191
76. Harrison RL, Jarvis DL (2007) Transforming lepidopteran insect cells for improved protein processing. Methods Mol Biol 388:341–356
77. Fernandes F, Teixeira AP, Carinhas N, Carrondo MJ, Alves PM (2013) Insect cells as a production platform of complex virus-like particles. Expert Rev Vaccines 12(2):225–236

78. Metz SW et al (2013) Effective chickungunya virus-like particle vaccine produced in insect cells. PLoS Negl Trop Dis 7:e2124
79. Smith GE et al (2013) Development of influenza H7N9 virus like particle (VLP) vaccine: homologous A/Anhui/1/2013 (H7N9) protection and heterologous A/chicken/Jalisco/CPA/2012 (H7N3) cross-protection in vaccinated mice challenged with H7N9 virus. Vaccine 31:4305–4313
80. Metz SW, Plijman GP (2011) Arbovirus vaccines: opportunities for the baculovirus-insect cell expression system. J Invertebr Pathol 107(Suppl):S16–S30
81. Behzadian F et al (2013) Baculoviral co-expression of HA, NA and M1 proteins of highly pathogenic H5N1 influenza virus in insect cells. Jundishapur J Microbiol 6(9):e7665
82. Condreay JP, Kost TA (2007) Baculovirus expression vectors for insect and mammalian cells. Curr Drug Targets 8(10):1126–1131
83. Ames RS, Kost TA, Condreay JP (2007) BacMam technology and its application to drug discovery. Expert Opin Drug Discov 2(12):1669–1681
84. Kost TA, Condreay JP, Ames RS (2010) Baculovirus gene delivery: a flexible assay development tool. Curr Gene Ther 10(3):168–173

Fundamentals of Expression in Mammalian Cells 14

Michael R. Dyson

Abstract

Expression of proteins in mammalian cells is a key technology important for many functional studies on human and higher eukaryotic genes. Studies include the mapping of protein interactions, solving protein structure by crystallization and X-ray diffraction or solution phase NMR and the generation of antibodies to enable a range of studies to be performed including protein detection *in vivo*. In addition the production of therapeutic proteins and antibodies, now a multi billion dollar industry, has driven major advances in cell line engineering for the production of grams per liter of active proteins and antibodies. Here the key factors that need to be considered for successful expression in HEK293 and CHO cells are reviewed including host cells, expression vector design, transient transfection methods, stable cell line generation and cultivation conditions.

Keywords

Antibody expression • Biologics production • Transient transfection • High-throughput expression

14.1 Introduction

For the expression of human and mammalian proteins, including antibodies, it is most appropriate to use a mammalian expression system with the chaperones, binding partners, secretion apparatus and post-translational modifications for correct protein folding. Successes have been achieved by the truncation of large complex multi domain containing proteins to individual domains and expression in *E. coli* [1] or the expression in yeast or insect cells. However, especially for secreted or membrane containing proteins, the non mammalian expression systems lack the appropriate machinery for authentic glycosylation [2–4]. This is an important consideration where the functional activity of a protein can change depending on the particular post-translation modification [5]. For example, mammalian cells synthesize complex glycans

M.R. Dyson (✉)
IONTAS Ltd, Babraham Research Campus,
Babraham, Cambridge CB22 3AT, UK
e-mail: mrd@iontas.co.uk

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_14

containing mannose, galactose, N-acetylglucosamine and sialic acid whereas insect cells mainly produce oligomannosidic and simpler paucimannosidic glycans and N-glycans [6] that contain α1,3-linked fucose residues, which are known to be allergenic.

Recent advances in heterologous protein expression in mammalian cells has allowed it's re-positioning from an activity that required specialist knowledge to one that is now a core activity as important to biology and biochemistry laboratories as molecular biology. Expression of proteins in mammalian cells was considered to be too expensive, required specialist equipment and staff and generally gave a poor yield for transient transfection or took too long to isolate a stable, high-producing cell line. However during the last 14 years a series of publications have emerged that have shown that, given the correct combination of cell line, transfection method, expression vector and cultivation conditions, excellent yields of protein can be achieved routinely and relatively quickly. The purpose of this chapter is not to provide an extensive review of all the advances in transient expression technologies as several excellent reviews have recently been published [7, 8]. Instead it is intended to summarize the current state of the art from the authors perspective, gained from a large research institute where a platform was set-up for high throughput protein expression of many different targets [9] to a biotechnology company where the multi-parallel expression of large numbers of antibody clones is important for screening purposes [10] and larger scale production is required for more detailed functional studies. This review will encompass cell-lines, transfection methods, cultivation conditions and stable cell line development and cell line engineering. For colleagues working in an academic laboratory or performing in house research in industry, the choice is driven purely by the optimal expression system that in many cases can be obtained as an "off the shelf" solution. However, in a commercial environment, additional costs may be incurred when licensing various expression technologies and this has driven the search for novel cell lines and expression vectors that do not have any "restricted use" limitations.

14.2 Host Cells

The main expression host cell lines are human embryonic kidney (HEK) 293 cells and Chinese hamster ovary (CHO) cells. Because of ease of cultivation and transfection, the HEK293 cell line is widely used in research laboratories. Originally the HEK293 cell line was established by the transformation of human embryonic kidney cells with sheared human adenovirus DNA, resulting in expression of the adenovirus E1A and E1B genes [11]. Subsequently the HEK293E [12] and HEK293T [13] cell lines were isolated by the integration of genes encoding the Epstein-Barr Virus nuclear antigen 1 (EBNA1) or simian virus 40 (SV40) large T antigen (LT) respectively. The function of EBNA1 and SV40 LT is to bind to their specific origins of replication (ori) and promote DNA replication by the recruitment of the DNA replication machinery. Therefore expression plasmids have been developed for transient gene expression (TGE) containing the SV40 or EBNA1 origins of replication [14] to aid episomal replication of plasmids and potentially increase protein expression yield. Although there are several reports to suggest that episomal plasmid replication in the HEK293T and HEK293E cell lines can lead to increased protein expression yield [15–17], this may be dependent on the transfection conditions. For example in one report the best expression yield was obtained with an expression vector lacking an origin of replication [18] and in our own laboratory we have compared the expression yield of antibodies in plasmids containing a SV40 origin of replication in suspension HEK293 Freestyle™, HEK293E and HEK293T cells and not observed a significant variation in yield (Dyson, M.R., unpublished data). There may, however, be an additional benefit of the presence of origin of replications which contain a nuclear import signal that aids in transport of the expression plasmid to the nucleus for transcription [19].

To enable high density cell growth for increased protein expression yield per unit volume of cell culture media and for simplified purification methods, suspension adapted cell lines able to grow in serum free conditions have an advantage. Examples of serum free suspension adapted cell lines include the HEK293 cell lines: 293 Freestyle™ [20] cells and Expi293F™ (Life Technologies) cells which contain no virally introduced elements. This laboratory has observed a greater than tenfold yield enhancement with the Expi293F™ cells compared with 293 Freestyle™ for antibody expression (data not shown). The HEK293-6E cell line, with integrated EBNA1 is suspension adapted to serum free media and available from NRC, Montreal, Canada under license [21]. For structural studies involving protein crystallization homogenous glycosylation is important and for this purpose a suspension adapted cell line lacking N-acetylglucosaminyltransferase-I activity (HEK293S GnTI^{-}, ATCC CRL-3022) has been developed [22]. It is possible to create a suspension adapted HEK293 cell line from adherent HEK293 cells using published protocols [23, 24].

CHO cells were originally isolated as spontaneously immortalized cells from primary Chinese hamster ovarian cultures [25]. CHO-K1 [26] and CHO-S [27] were derived from the original CHO cell line with the latter being adapted to suspension culture. A CHO-K1 derivative cell line has also been constructed with integrated EBNA1 and glutamine synthase genes for enhanced transgene expression [28]. A CHO cell line adapted for suspension growth in serum free conditions, also designated CHO-S cells, is available from Life Technologies. In addition the CHO-DG44 cell line [29] was derived by gamma irradiation of the original CHO cell line followed by a screen for the absence of dihydrofolate reductase (DHFR) activity. CHO-DG44 cells enable the selection of stable integrated DHFR cassettes also encoding a gene of interest (GOI) in media lacking hypoxanthine and thymidine. CHO-DG44 cells are widely used in the biopharmaceutical industry to generate stable cell lines for therapeutic antibody production.

Alternative hosts have been developed [7, 30], but for the majority of expression projects HEK293 or CHO cell lines and their derivatives are sufficient for the vast majority of expression projects. The only caveat to this is if one desired to screen libraries directly for function in which case it is better to choose a host cell relevant to the particular function to be tested. For example antibody libraries were recently integrated into the Rosa26 locus of mouse stem cells so that one antibody was integrated per cell. Individual clones were then screened for their ability to retain a pluripotent phenotype under neuronal differentiation conditions [31]. In this way novel blockers of the FGFR signaling pathway were identified.

14.3 Vector Design

The key elements to consider for a successful expression vector are choice of promoter, the presence of an intron in the 5′ UTR, signal peptide if the protein is to be secreted, polyadenylation site and 3′ UTR. The human cytomegalovirus (hCMV) major immediate early (mIE) promoter/enhancer is a popular choice that works well in both HEK293 and CHO cells. An antibody expression vector (pXL-G^{HEK}) has been described capable of yielding high antibody expression yields under optimized culture conditions [18]. The vector consisted of separate light and heavy chain expression vectors which were co-transfected. Each vector contained a CMV promoter, SV40 intron upstream of the Kozak consensus sequence, kappa light chain signal peptide, woodchuck post-transcriptional regulatory element (WPRE) and bovine growth hormone (BGH) polyadenylation sequence. An alternative expression plasmid that has been used to express single chain variable fragment (scFv) Fc fusion proteins is pBIOCAM5 [31, 32] which is available from Addgene (https://www.addgene.org/39344/). This plasmid consists of a backbone originally derived from pCMV/myc/ER (Lifetech), a CMV promoter from pCEP4 (Lifetech) and a cassette from pAdCMV5 [33] consisting of a tripartite leader sequence for enhanced translation efficiency, adenovirus major late promoter enhancer and an intron for efficient mRNA export. In addition pBIOCAM5 contains a signal peptide encoding sequence with an embedded intron originally from the human antibody kappa constant light chain leader sequence. pBIOCAM5 routinely gives an expression yield of between 20 and 30 mg/L of secreted scFv-Fc by polyethyleneimine (PEI) transfection of HEK293F cells under standard conditions.

Antibody expression plasmids can also be constructed where the heavy and light chains are expressed on the same plasmid. Here the vector can be bicistronic with two separate promoters driving heavy and light chain expression or be mono-cistronic with an internal ribosome entry (IRES) site [34]. Alternatively the two polypeptides can be separated by a picornavirus 2A self-processing peptide [35]. For antibody expression cassettes, or any expression cassette encoding components of a protein complex it is important to consider the relative strengths of the promoters driving expression. For example if a 1:1 ratio of two components are desired in the final complex then it is often the safest strategy to drive expression with two identical promoters. Alternatively if an excess of one component compared to a second component in the final complex is necessary or if component A acts as a chaperone for the correct folding of component B then it may be advisable to drive the expression of component A with a stronger promoter than component B. For example the CMV promoter is known to be stronger than the elongation factor 1 (EF1) promoter [18]. Finally when expressing receptor ectodomains a successful strategy is often to express with the native signal peptide of that protein. If a protein normally exists as a heterodimer then soluble expression can be aided by co-expression of it's binding partner or truncation to express individual protein domains [36]. Finally the majority of promoters used for expression are constitutive. However for the expression of toxic genes there may be an advantage in using an inducible promoter such as the "Tet-on" promoter system [37].

14.4 DNA Transfection

Although efficient transfection can be achieved with cationic lipid formulations for both HEK293 and CHO cells using reagents such as FuGene HD (Roche) or Lipofectamine 2000 (Lifetech) [7], these methods are not scalable due to their expense. Calcium phosphate has been used successfully to transfect CHO and HEK293 cells [38], but requires the presence of serum which is not compatible with the many serum free media compositions commonly used. PEI is commonly used as a transfection reagent for both HEK293 [8] and CHO [39] due to its low cost and high transfection efficiency. Linear PEI is most commonly used and can be purchased from Polysciences Inc. and prepared as described previously [40] or as a pre-prepared solution from PolyPlus. A 2:1 ratio of PEI to DNA is commonly used for HEK293 transfection whereas higher ratios are usually used for CHO transfection [39, 41]. Figure 14.1 shows SDS PAGE separation of purified human IgG2 expressed by PEI transfection of HEK293F cells. PEI transfection efficiency can alter according to the cultivation media used [7] and this may be due the presence of heparin sulfate, dextran or iron (III) citrate [42]. The highest reported transient expression yields so far have been achieved by PEI transfection. For example Backliwal et al. [18] developed an optimized protocol that involved high cell density transfection, co-

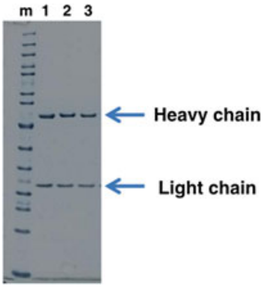


Fig. 14.1 SDS PAGE of purified antibodies. Purified anti-Notch1 IgG antibodies [32] were separated by SDS-PAGE and stained with Coomassie blue stain. The marker lane (m) is PAGE ruler ladder (Fermentas, 26614). The antibody heavy and light chains are indicated by *arrows*

transfection with cell cycle regulators p18 and p21 and Fibroblast growth factor. Valproic acid, an inhibitor of histone deacetylase, was added 3 h post transfection and the cells maintained at a high density of four million cells per ml of culture. This achieved an antibody expression yield of 1 g/L. Daramola et al. [28] exceeded this expression yield by PEI transient transfection of a suspension adapted CHO-K1 cell line stably expressing EBNA-1 and glutamine synthase (GS) in an optimized protocol. Recently hyper branched polylysine has been suggested as a biodegradable alternative transfection reagent to PEI [43]. Also flow electroporation methods have been described for transfection of CHO-S cells at the liter scale achieving antibody expression yields of greater than 1 g per liter [44].

14.5 Methods for Stable Cell Line Production/Transposase Mediated Integration

The majority of small to medium scale expression experiments can be performed by transient gene expression protocols. However if multi-milligram quantities of protein are required, it can be tedious to produce the large amounts of plasmid DNA required for large scale transfection. Here it can be more convenient to isolate a stable cell line or stable pooled cell line expressing the gene of interest [45]. Traditionally stable cell line development has involved transfection with linear DNA encoding an expression cassette and a selectable marker. Random genome integration is mediated by non homologous end joining processes [46]. Recently transposase mediated gene integration has enable the rapid generation of pools of stable cell lines expressing a gene of interest [47, 48] and in our hands this has resulted in a greater than tenfold yield improvement after 3–4 weeks of selection compared with standard transient transfection of HEK293F cells. Transposase mediated gene integration results on average in three to four copies of an expression cassette being integrated. If a single copy gene integration into a specific locus is desired alternative nuclease mediated integration methods are available including the TALE nuclease system [49] or CRISPR/Cas9 [50].

14.6 Cultivation Conditions

A successful expression experiment in mammalian cells is crucially dependent of the health of the cells in terms of growth rate and cell viability. Each cell line has been adapted to growth in a particular growth media [7]. Suspension adapted cell lines are routinely grown in conical flasks ranging in volume from 125 ml to 2 L sourced from companies such as Corning (see Fig. 14.2). Optimum growth flasks from Thomson Instrument Company, ranging in size from 125 ml to 5 L are superior to standard flasks in terms of their ability to maintain cell viability during the course of an expression experiment. TubeSpin® 50 ml or 600 ml bioreactors (TPP) have also been used successfully. The growth of suspension adapted HEK293 and CHO cells is generally performed in humidified CO_2 shake incubators, but can also be grown in closed flasks pre-gassed with CO_2. Suspension adapted HEK293 or CHO cells can also be grown in 24-well blocks in 2 ml or 4 ml cultures [10, 51] or 96-well plates [52] to enable multi-parallel expression.

14.7 Conclusions

To summarize all the factors listed above including cell line, vector design, transfection and cultivation methods need to be considered for a successful expression project in mammalian cells. The end objective in terms of the number of clones to be expressed and yield desired also need to be considered and this can help to decide on transient expression versus stable cell line generation. The recent developments in this field have now opened the door for the routine expression of targets previously considered to be difficult and enable high throughput expression experiments including antibody screening [53] previously only possible by expression in *E. coli* or yeast.

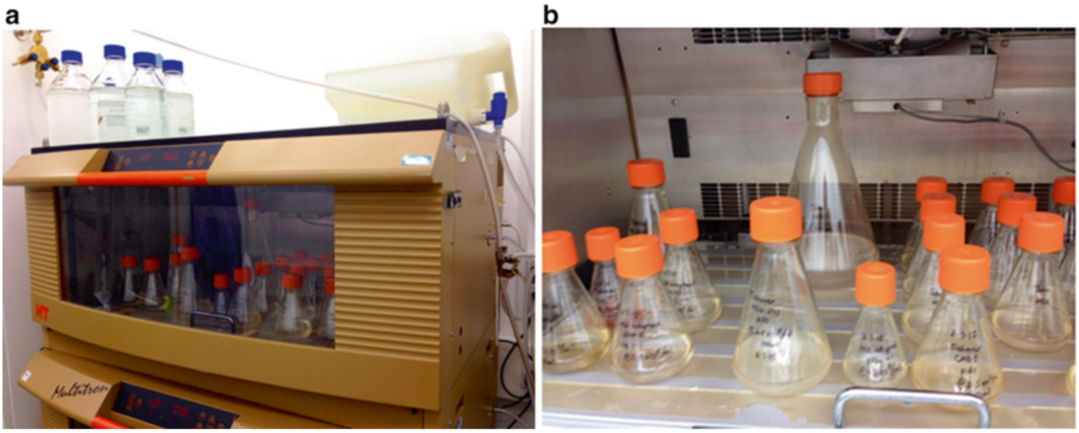


Fig. 14.2 Cultivation of HEK293 and CHO suspension cells. Suspension HEK293 and CHO cells are cultivated in a humidified CO_2 shake incubator as supplied by Infors (**a**) or Kuhner. Cells are grown in sterile Erlenmeyer flasks with vented caps (**b**) so that the volume of media and cells is no more than 25 % of the flask capacity with a shake speed of 130 rpm (25 mm orbital throw), 5 % CO_2 and 75 % humidity

References

1. Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J (2004) Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression. BMC Biotechnol 4:32
2. Marchal I, Jarvis DL, Cacan R, Verbert A (2001) Glycoproteins from insect cells: sialylated or not? Biol Chem 382(2):151–159
3. Byrne B, Donohoe GG, O'Kennedy R (2007) Sialic acids: carbohydrate moieties that influence the biological and physical properties of biopharmaceutical proteins and living cells. Drug Discov Today 12(7–8):319–326
4. Nallet S, Fornelli L, Schmitt S, Parra J, Baldi L, Tsybin YO, Wurm FM (2012) Glycan variability on a recombinant IgG antibody transiently produced in HEK-293E cells. New Biotechnol 29(4):471–476
5. Jefferis R (2009) Recombinant antibody therapeutics: the impact of glycosylation on mechanisms of action. Trends Pharmacol Sci 30(7):356–362
6. Hossler P, Khattak SF, Li ZJ (2009) Optimal and consistent protein glycosylation in mammalian cell culture. Glycobiology 19(9):936–949
7. Geisse S, Voedisch B (2012) Transient expression technologies: past, present, and future. In: Voynov V, Caravella JA (eds) Therapeutic proteins, vol 899. Humana Press, New York, pp 203–219
8. Hacker DL, Kiseljak D, Rajendra Y, Thurnheer S, Baldi L, Wurm FM (2013) Polyethyleneimine-based transient gene expression processes for suspension-adapted HEK-293E and CHO-DG44 cells. Protein Expr Purif 92(1):67–76
9. Schofield DJ, Pope AR, Clementel V, Buckell J, Chapple S, Clarke KF, Conquer JS, Crofts AM, Crowther SR, Dyson MR et al (2007) Application of phage display to high throughput antibody generation and characterization. Genome Biol 8(11):R254
10. Chapple SD, Dyson MR (2014) Expression screening in mammalian suspension cells. Methods Mol Biol 1091:143–149
11. Graham FL, Smiley J, Russell WC, Nairn R (1977) Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J Gen Virol 36(1):59–74
12. Yates JL, Warren N, Sugden B (1985) Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells. Nature 313(6005):812–815
13. Rio DC, Clark SG, Tjian R (1985) A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. Science 227(4682):23–28
14. Van Craenenbroeck K, Vanhoenacker P, Haegeman G (2000) Episomal vectors for gene expression in mammalian cells. Eur J Biochem/FEBS 267(18):5665–5678
15. Cachianes G, Ho C, Weber RF, Williams SR, Goeddel DV, Leung DW (1993) Epstein-Barr virus-derived vectors for transient and stable expression of recombinant proteins. Biotechniques 15(2):255–259
16. Parham JH, Kost T, Hutchins JT (2001) Effects of pCIneo and pCEP4 expression vectors on transient and stable protein production in human and simian cell lines. Cytotechnology 35(3):181–187
17. Durocher Y, Perret S, Kamen A (2002) High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. Nucleic Acids Res 30(2), e9
18. Backliwal G, Hildinger M, Chenuet S, Wulhfard S, De Jesus M, Wurm FM (2008) Rational vector design and multi-pathway modulation of HEK 293E cells

yield recombinant antibody titers exceeding 1g/l by transient transfection under serum-free conditions. Nucleic Acids Res 36(15), e96

19. Dean DA (1997) Import of plasmid DNA into the nucleus is sequence specific. Exp Cell Res 230(2):293–302
20. Liu C, Dalby B, Chen W, Kilzer JM, Chiou HC (2008) Transient transfection factors for high-level recombinant protein production in suspension cultured mammalian cells. Mol Biotechnol 39(2):141–153
21. Raymond C, Tom R, Perret S, Moussouami P, L'Abbé D, St-Laurent G, Durocher Y (2011) A simplified polyethylenimine-mediated transfection process for large-scale and high-throughput applications. Methods 55(1):44–51
22. Reeves PJ, Callewaert N, Contreras R, Khorana HG (2002) Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. Proc Natl Acad Sci U S A 99(21):13419–13424
23. Tsao YS, Condon R, Schaefer E, Lio P, Liu Z (2001) Development and improvement of a serum-free suspension process for the production of recombinant adenoviral vectors using HEK293 cells. Cytotechnology 37(3):189–198
24. Li L, Qin J, Feng Q, Tang H, Liu R, Xu L, Chen Z (2011) Heparin promotes suspension adaptation process of CHO-TS28 cells by eliminating cell aggregation. Mol Biotechnol 47(1):9–17
25. Puck TT, Cieciura SJ, Robinson A (1958) Genetics of somatic mammalian cells: III. Long term cultivation of euploid cells from human and animal subjects. J Exp Med 108(6):945–956
26. Kao FT, Puck TT (1968) Genetics of somatic mammalian cells, VII. Induction and isolation of nutritional mutants in Chinese hamster cells. Proc Natl Acad Sci U S A 60(4):1275–1281
27. Gottesman MM (1987) Chinese hamster ovary cells. In: Gottesman M (ed) Methods in enzymology, vol 151. Academic Press Inc, San Diego, pp 3–8
28. Daramola O, Stevenson J, Dean G, Hatton D, Pettman G, Holmes W, Field R (2014) A high-yielding CHO transient system: coexpression of genes encoding EBNA-1 and GS enhances transient protein expression. Biotechnol Prog 30(1):132–141
29. Urlaub G, Käs E, Carothers AM, Chasin LA (1983) Deletion of the diploid dihydrofolate reductase locus from cultured mammalian cells. Cell 33(2):405–412
30. Almo SC, Love JD (2014) Better and faster: improvements and optimization for mammalian recombinant protein production. Curr Opin Struct Biol 26(1):39–43
31. Melidoni AN, Dyson MR, Wormald S, McCafferty J (2013) Selecting antagonistic antibodies that control differentiation through inducible expression in embryonic stem cells. Proc Natl Acad Sci U S A 110(44):17802–17807
32. Falk R, Falk A, Dyson MR, Melidoni AN, Parthiban K, Young JL, Roake W, McCafferty J (2012) Generation of anti-Notch antibodies and their application in blocking Notch signalling in neural stem cells. Methods 58(1):69–78
33. Massie B, Mosser DD, Koutroumanis M, Vitte-Mony I, Lamoureux L, Couture F, Paquet L, Guilbault C, Dionne J, Chahla D et al (1998) New adenovirus vectors for protein production and gene transfer. Cytotechnology 28(1–3):53–64
34. Underhill MF, Smales CM, Naylor LH, Birch JR, James DC (2007) Transient gene expression levels from multigene expression vectors. Biotechnol Prog 23(2):435–443
35. Fang J, Qian JJ, Yi S, Harding TC, Tu GH, VanRoey M, Jooss K (2005) Stable antibody expression at therapeutic levels using the 2A peptide. Nat Biotechnol 23(5):584–590
36. Dyson MR (2010) Selection of soluble protein expression constructs: the experimental determination of protein domain boundaries. Biochem Soc Trans 38(4):908–913
37. Gossen M, Bujard H (1992) Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc Natl Acad Sci U S A 89(12):5547–5551
38. Jordan M, Schallhorn A, Wurm FM (1996) Transfecting mammalian cells: optimization of critical parameters affecting calcium-phosphate precipitate formation. Nucleic Acids Res 24(4):596–601
39. Rajendra Y, Kiseljak D, Baldi L, Hacker DL, Wurm FM (2011) A simple high-yielding process for transient gene expression in CHO cells. J Biotechnol 153(1–2):22–26
40. Tom R, Bisson L, Durocher Y (2007) Transient expression in HEK293-EBNA1 cells. In: Dyson MR, Durocher Y (eds) Expression systems. Scion, Bloxham, pp 203–223
41. Ye J, Kober V, Tellers M, Naji Z, Salmon P, Markusen JF (2009) High-level protein expression in scalable CHO transient transfection. Biotechnol Bioeng 103(3):542–551
42. Eberhardy SR, Radzniak L, Liu Z (2009) Iron (III) citrate inhibits polyethylenimine-mediated transient transfection of Chinese hamster ovary cells in serum-free medium. Cytotechnology 60(1–3):1–9
43. Kadlecova Z, Rajendra Y, Matasci M, Hacker D, Baldi L, Wurm FM, Klok HA (2012) Hyperbranched polylysine: a versatile, biodegradable transfection agent for the production of recombinant proteins by transient gene expression and the transfection of primary cells. Macromol Biosci 12(6):794–804
44. Steger K, Brady J, Wang W, Duskin M, Donato K, Peshwa M (2015) CHO-S antibody titers >1 gram/liter using flow electroporation-mediated transient gene expression followed by rapid migration to high-yield stable cell lines. J Biomol Screen 20(4):545–551
45. Büssow K (2015) Stable mammalian producer cell lines for structural biology. Curr Opin Struct Biol 32:81–90

46. McVey M, Lee SE (2008) MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet: TIG 24(11):529–538
47. Matasci M, Baldi L, Hacker DL, Wurm FM (2011) The PiggyBac transposon enhances the frequency of CHO stable cell line generation and yields recombinant lines with superior productivity and stability. Biotechnol Bioeng 108(9):2141–2150
48. Balasubramanian S, Matasci M, Kadlecova Z, Baldi L, Hacker DL, Wurm FM (2015) Rapid recombinant protein production from piggyBac transposon-mediated stable CHO cell pools. J Biotechnol 200:61–69
49. Maresca M, Lin VG, Guo N, Yang Y (2013) Obligate Ligation-Gated Recombination (ObLiGaRe): custom-designed nuclease-mediated targeted integration through nonhomologous end joining. Genome Res 23(3):539–546
50. Liang X, Potter J, Kumar S, Zou Y, Quintanilla R, Sridharan M, Carte J, Chen W, Roark N, Ranganathan S et al (2015) Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. J Biotechnol 208:44–53
51. Chapple SD, Crofts AM, Shadbolt SP, McCafferty J, Dyson MR (2006) Multiplexed expression and screening for recombinant protein production in mammalian cells. BMC Biotechnol 6:49
52. Vink T, Oudshoorn-Dickmann M, Roza M, Reitsma J-J, de Jong RN (2014) A simple, robust and highly efficient transient expression system for producing antibodies. Methods 65(1):5–10
53. Dyson MR, Zheng Y, Zhang C, Colwill K, Pershad K, Kay BK, Pawson T, McCafferty J (2011) Mapping protein interactions by combining antibody affinity maturation and mass spectrometry. Anal Biochem 417(1):25–35

15 Assembling Multi-subunit Complexes Using Mammalian Expression

Bahar Baser and Joop van den Heuvel

Abstract

In this chapter conventional and emerging new technologies for the production of complex biologics in mammalian expression systems are summarized. The essential features of the most relevant methods to generate stable production cell lines for the expression of recombinant multi-protein complexes are described. Especially the promising multiple targeted integration strategy by Flp or CRISPR/Cas9 mediated recombination and their future impact on multi-protein expression are highlighted.

Keywords

Targeted integration • Stable cell line • Flp/*FRT* recombination • Multi RMCE • MultiMam

15.1 Introduction

Most proteins do not act as singular entities but are part of complex assemblies composed of several subunits. Their involvement in many cellular processes makes it imperative to elucidate their structural composition and function. Therefore an understanding of protein-protein interactions present within the complex or with other complexes is required to determine the functional position of a protein complex within the "molecular sociology" of a cellular network [1]. Structural biology plays a major role in the elucidation of protein-protein interactions at atomic resolution. However, the requirement of protein in sufficient quantity and quality is a major bottleneck. Recombinant protein expression is required to obtain proteins that are often just expressed in low amounts in the endogenous host cell [2]. Unfortunately not every protein subunit can be expressed in soluble form by itself. Proteins often depend on their partner within multi-protein complexes. Co-expression of protein subunits may allow proper folding, assembly and finally soluble expression of multi-subunit complexes. Moreover, the recombinant co-expression of protein subunits can permit exact adjustment of expression ratios for each individual subunit to optimize for functional over-expression [3, 4].

B. Baser • J. van den Heuvel (✉)
Department of Structure and Function of Proteins,
Helmholtz Centre for Infection Research,
Braunschweig, Germany
e-mail: joop.vandenheuvel@helmholtz-hzi.de

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_15

The choice of the expression host may have a major impact on successful expression of a target protein or protein complex. Prokaryotic expression systems, particularly *Escherichia coli*, are well established in most laboratories. While several *E. coli* systems are available for co-expression of proteins [5–7], eukaryotic multi-subunit protein targets often require more complex expression hosts. Eukaryotic host cells are able to generated posttranslational modifications (PTMs) and allow proper processing of eukaryotic recombinant proteins. Moreover, they provide the complete secretory pathway necessary for proper folding and secretion of complex proteins [8]. Particularly mammalian expression hosts provide the most native environment for human proteins. This chapter will describe current strategies for the expression of multi-subunit complexes in mammalian expression hosts.

15.2 Conventional Strategies for Recombinant Protein Production in Mammalian Cells

Conventional strategies for co-expression of proteins in mammalian expression hosts include transient as well as stable expression strategies. The methods range from approaches using viral vectors, recombinant plasmids or genetic recombination based technologies. The co-expression of proteins can be achieved using multiple vectors carrying each a single expression cassette (Fig. 15.1a) or as a single vector carrying all required open reading frames (ORFs) (Fig. 15.1b–d). Using the single expression vector strategy there are three options. Either, each ORF can be transcribed in a separate transcription unit comprising its own promoter and termination signal. Or alternatively, all ORFs can be transcribed in a single transcription unit, which might contain either a polycistronic or a monocistronic expression unit [9] (Fig. 15.1c, d). Multiple as well as single expression vectors have been successfully used for the co-expression of protein subunits [9]. However, due to different individual transfection efficiencies for each vector using the multi vector approach, not all genes are necessarily present in each cell after transient transfection. In contrast, the entry of a single vector will assure that all expression units for the desired proteins or protein subunits are expressed within this transfected cell. Additionally, different relative ratios of expression of the required subunits of the protein complex can be adjusted by the use of gene copy number, specific promoters [10], polyadenylation signals [11] or internal ribosome entry sites (IRES) [4, 12] of varying strength. Special care has to be taken to avoid interference of the cassettes by close proximity. This may influence transcriptional regulation, induce suppression or result in silencing effects [13, 14].

15.2.1 Transient Expression in Mammalian Hosts

In the early years of transient protein production mainly engineered lenti- or adenoviral vectors were utilized for strong recombinant expression. However, in the last decade plasmid based transfection methods greatly improved in transfection efficiency and productivity. Currently, plasmid-based transient transfection is the preferred method for fast gene delivery into mammalian cells, because cumbersome steps such as virus expansion and storage can be eliminated [15]. Compared to the generation of stable mammalian cell lines, transient expression offers a fast way to screen for expressible constructs and produce recombinant protein. Today, transient transfection of human embryonic kidney 293 (HEK293) derived cell lines including HEK293T [16], HEK293-EBNA [17] and HEK293-6E [18] is the most commonly and best established method used for transient mammalian protein production. Adherent HEK293 cell lines can be used for construct screening and protein production in a small scale [19]. Likewise, suspension adapted HEK293 cell lines are used for transient screening in small scale multiwell formats [20] as well

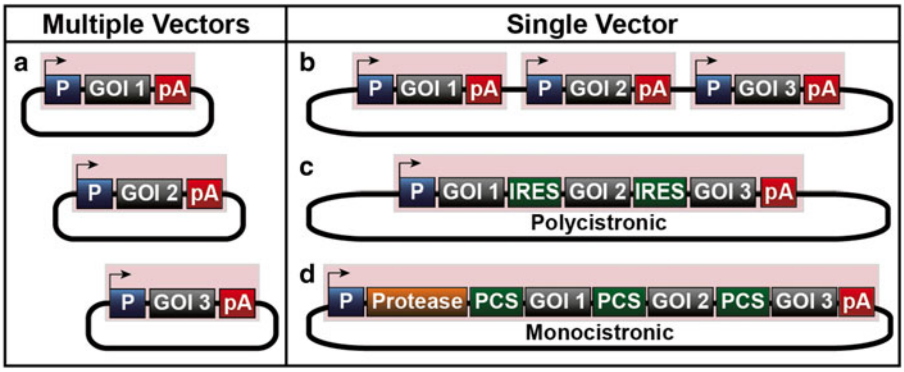

Multiple Vectors
Single Vector
a
P GOI 1 pA
P GOI 2 pA
P GOI 3 pA
b
P GOI 1 pA P GOI 2 pA P GOI 3 pA
c
P GOI 1 IRES GOI 2 IRES GOI 3 pA
Polycistronic
d
P Protease PCS GOI 1 PCS GOI 2 PCS GOI 3 pA
Monocistronic

Fig. 15.1 Expression vectors for multi-gene expression. Schematic overview for multi-gene expression exemplified for three genes of interest (GOI). (**a**) Multiple expression vectors, each transcribing one gene, can be used for co-expression. These can either be integrated in a sequential fashion into the host genome to generate stable cell lines or can be used for protein production through simultaneous transient co-transfection. (**b**) The use of a single expression vector, comprising several expression cassettes allows balanced expression of all desired genes in the same cell. (**c**) A polycistronic expression vector enables the individual protein synthesis of each GOI from one mRNA transcript carrying multiple translational units separated by internal ribosome entry sites. (**d**) A monocistronic expression vector (single translation unit) will result in a polyprotein, which will subsequently be processed by a specific protease to generate the individual protein products. *Abbreviation*'*s*: *P* promoter, *GOI* gene of interest, *pA* transcription termination and polyadenylation signal, *IRES* internal ribosome entry sites, *PCS* protease cleavage site (Adapted from [9])

as for large scale expression in Wave™ bioreactors [21, 22]. The EBNA expressing cell lines allow the episomal amplification of plasmids containing the corresponding origin of replication (oriP). Due to the prolonged retention of plasmids within the cell, transient yields for recombinant protein can be substantially increased. For special applications like structural biology often the glycosylation mutant cell line HEK293-*GnTI(-)* is used [22, 23]. The secreted protein product of the glycosylation mutant cell line can be easily deglycosylated by Endo H treatment to generate homogenous material for crystallization and 3D structural analysis.

Alternatively, laboratories with established baculoviral infected insect cell platforms can also use the baculovirus based expression in mammalian cells (BacMam). As baculoviruses are not able to replicate in mammalian cells the BacMam system has a favorable biosafety profile over other viral methods. Moreover, the integration of large expression cassettes (up to 38 kb) is feasible which makes the BacMam system useful for multi-gene expression [24–28].

15.2.2 Stable Expression in Mammalian Cells

The generation of stable production cell lines offers certain advantages over transient protein production. Batch-to-batch variations in quality and yield will be mostly eliminated using stable production cell lines and therefore, these cell lines offer a more reliable process to reproducibly scale up protein production. However, standard methods for stable cell line generation are often very time consuming because of the random introduction of expression cassettes into the genome. This results in unpredictable expression patterns due to differences in the site of integration (position-effect) and the number of transgene integrations (gene dose). Extensive screening for high producer cell lines is required which is both labor- and cost-intensive [29]. Chinese hamster ovary (CHO) cell lines are the favored host for the production of therapeutic proteins. Different selection methods such as mammalian selection markers, like the dihydrofolate reductase (DHFR) system or the glutamine synthase (GS) system are commonly utilized [29]. The serial dilution technique is the predominant method for the isolation of the final clonal production cell line. Alternative methods to generate stable cell lines are using lentiviral vectors, which favor the integration into transcriptionally active sites [30–32].

15.3 Tandem Recombineering for Multi-protein Co-expression

To simultaneously co-express proteins or protein subunits within a host, it is necessary to introduce all required recombinant genes into the same cell. Tandem recombineering offers a fast way to assemble multigene expression vectors using a donor-acceptor strategy via Cre/*lox* recombination. Initially introduced as MultiBac system for multi-gene expression in insect cells [33, 34] and later on as ACEMBL in *E. coli* [35] tandem recombineering was also adapted for use in mammalian expression hosts [36, 37].

The MultiMam system utilizes an array of donor and acceptor vectors with multiplication modules consisting of a homing endonuclease and a *Bst*XI site. These flank a multiple cloning site (MCS) that also comprises a promoter (P) and a polyadenylation signal (pA). Upon insertion of a gene of interest (GOI) into the MCS of either an acceptor or donor vector the entire expression cassette (P-GOI-pA) can be excised via the homing endonuclease and *Bst*XI sites. One or more excised expression cassette can be introduced into the multiplication module of another compatible donor or acceptor vector via the *Bst*XI restriction site. Upon insertion of one or more recombinant genes into an acceptor or donor vector, Cre/*lox* recombination is used to fuse the donor vectors with an acceptor vector to obtain a multi-gene expression plasmid. As donor vectors comprise a conditional origin of replication (R6Kγ) they can only propagate in *E. coli* that express the *pir* gene. Therefore only fusion with an acceptor vector will allow propagation in *pir*(−) negative *E. coli* [37, 38] (Fig. 15.2).

Similarly the MultiLable system uses tandem recombineering to assemble a single multi-gene expression vector that contains all genes of interest in separate expression cassettes. In contrast to the MultiMam system, donor and acceptor vectors are assembled in a single step from eight modular fragments selected according to the desired properties using pre-defined cohesive ends. The fragments include a *lox*P site for Cre-mediated recombination of donor and acceptor vectors, a mammalian promoter of choice, a fluorescent protein or a tag for each vector, a linker region that serves as a placeholder for a GOI and finally a polyA signal. A special region can be designed to accommodate specific requirements such as mammalian selection markers or homing endonuclease sites. Upon *in vitro* assembly of the donor vectors and acceptor vector using Cre-mediated recombination the multigene vector is propagated in *pir*(−) *E. coli* using the same principle as described for the MultiMam system. Only fusion with the acceptor vector will allow propagation in *pir*(−) *E. coli* [36].

Tandem recombineering offers a fast and flexible technology to generate multi-gene expression vectors for the transient or stable production of recombinant proteins. In contrast to conventional single vector strategies that rely on restriction based cloning alone, optimization of the construct can be done for each individual component by exchanging the specific gene by disassembly/assembly with Cre recombinase using another donor vector with a modified or optimized version of the challenging gene.

15.4 Targeted Integration for Multi-protein Expression in Stable Cell Lines

Random integration of transgenes into the genome of the expression host is still the most utilized strategy to obtain stable high producer cell lines for protein production. To improve the control over the site of integration, alternative approaches that utilize site-specific targeted integration for the generation of stable cell lines are available.

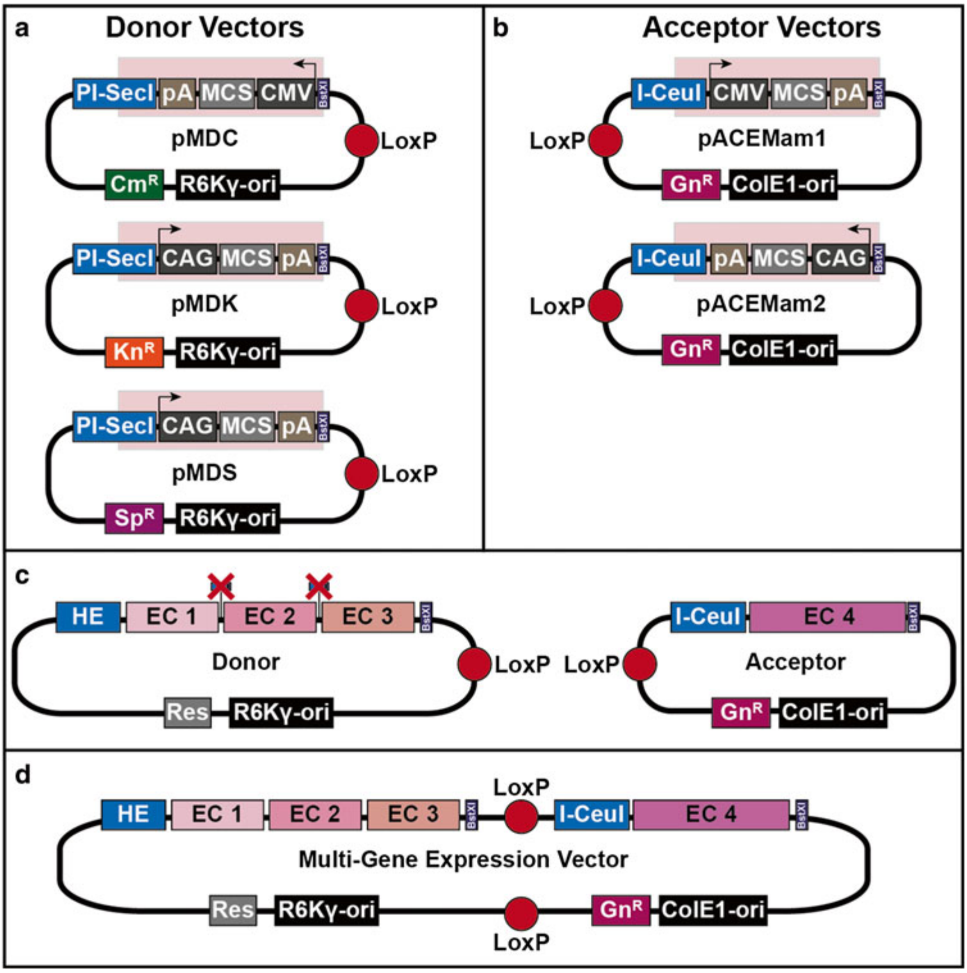


Fig. 15.2 MultiMam system. The MultiMam system utilizes an array of (**a**) donor and (**b**) acceptor vectors. Both comprise multiplication modules that allow the introduction of several GOIs into the multiple cloning site (MCS). The entire expression cassette (*rose color*) can be excised via homing endonuclease (HE, PI-SecI or I-CeuI) and *Bst*XI sites. (**c**) Upon excision, the expression cassette (EC) can be introduced into the multiplication module of a compatible vector via the *Bst*XI restriction site. The *Bst*XI site is designed to match the 3′-overhang of the respective homing endonuclease sites of the excised expression cassette. The introduction of an expression cassette via the *Bst*XI site destroys the homing endonuclease site but retains the *Bst*XI site for the introduction of additional expression cassettes. (**d**) After the introduction of one or more expression cassettes into the multiplication module of an acceptor or donor vector Cre/*lox* recombination is used to combine one or more donor vectors with an acceptor vector. Only donor vectors that successfully integrated in the acceptor vector will be able to propagated in *pir*(−) *E. coli* strains under appropriate antibiotic pressure. *HE* homing endonuclease (PI-SecI or I-CeuI), *pA* polyadenylation signal, *MCS* multiple cloning site, *CMV* cytomegalovirus promoter, *CAG* CAG promoter, *Cm*^{R} chloramphenicol, *Kn*^{R} kanamycin, *Sp*^{R} spectinomycin, *Gn*^{R} gentamycin, *Res* resistance marker, *EC* expression cassette (Adapted from [37, 38])

Transposons such as *Sleeping Beauty* and *piggyBac* are mobile genetic elements that target naturally occurring sites within mammalian genomes. While the integration of multiple transgenes is possible, neither their location nor the exact number of transgene integrations can be controlled. Moreover, integration may occur in transcriptionally active or inactive sites [39–41]. Due to its higher affinity for transcriptionally active sites [40, 42], *piggyBac* is preferred for recombinant protein production over *Sleeping Beauty*. So far however, *piggyBac* was only used for the quick generation of polyclonal pools producing extracellular receptor-Fc fusion proteins or ER-resident proteins in mammalian cells [43–45]. Nevertheless, the capability of *piggyBac* for co-expression of several proteins or protein subunits was also shown for cell reprogramming approaches [46, 47] and functional assays in mammalian cells [48].

Efficient generation of isogenic producer cell lines is favored for obtaining long-term reproducible protein expression method. Therefore, site-specific recombination technologies that target previously tagged and validated chromosomal loci were established for integration of transgenes into mammalian genomes [49, 50]. The use of site specific recombinases in tag-and-target (targeted integration) [49, 51] and tag-and-exchange (targeted replacement) strategies [52] currently are the method of choice for the generation of stable mammalian cell lines. But customized nucleases, such as the CRISPR/Cas9 system, emerge as alternative systems for the targeted integration of transgenes into mammalian genomes as whole genome sequences of expression hosts become available [53].

The current state for the targeted integration via site specific recombinases and customized nucleases in mammalian cells will be described in the next sections.

15.4.1 Engineered Nucleases for Sequence Specific Genome Editing

Customized nucleases such as Zinc-finger nucleases (ZFNs) [54, 55], transcription activator-like effector nucleases (TALENs) and clustered, regulatory interspaced, short palindromic repeats (CRISPR) associated protein 9 (Cas9) nucleases [56, 57] are frequently used for genomic modifications. Particularly ZFNs have been used for *in vivo* gene knockouts, replacements or repair in gene therapy as an alternative to gene silencing [53, 58]. They all induce site-specific DNA double strand breaks (DSBs). Zinc-finger nucleases and TALENs recognize DNA sequences through an array of linked protein–DNA binding domains to guide their chimeric nuclease to a specific chromosomal location. The CRISPR/Cas9 system on the contrary uses base-pairing of CRISPR-RNA (crRNA) which contains the homologous protospacer region recognizing a specific target sequence and the trans-activating crRNA (tracrRNA) which recruits the Cas9 nuclease to a desired locus, as described in [58]. Alternatively, a fusion of both crRNA and tracrRNA units, referred to as guideRNA (gRNA), can be used for chromosomal targeting of Cas9 nuclease (Fig. 15.3a). GuideRNAs are much easier to design and can be produced more cost-efficiently than their protein counterparts in ZNF and TALEN systems. Therefore, the CRISPR/Cas9 system is better suitable for the simultaneous introduction of several gRNAs at different genomic sites (multiplexing strategies) [58, 59].

It was shown that the CRISPR/Cas9 system could be used for the targeted integration of genes into CHO cells to generate populations with stable transgene expression. While the integration into different sites is possible, the number of integrations is not foreseeable [60]. Moreover, potential off-target sites are difficult to predict. Efforts to reduce off-target effects were examined. These included the use of lower gRNA concentrations, truncated or elongated gRNA constructs as well as the use of paired nickases but neither offered a holistic solution [58].

However, potential off-target mutations are not the major limitation for recombinant protein production. DSBs are repaired through two different pathways, either the predominant non-homologous end joining (NHEJ) pathway or the homology directed repair (HDR) pathway (Fig. 15.3). NHEJ is error-prone and often introduces insertions or deletions (indels) causing frameshifts within an ORF. While this is acceptable for knock-out mutations, the integration and in frame fusion of transgenes does require HDR for precise insertion. Though the use of nickases, generating single strand breaks, showed some success to shift the balance towards the HDR pathway, further solutions to circumvent this bottleneck are required. Alternatively, the use of a bait sequence adjacent to the integration DNA

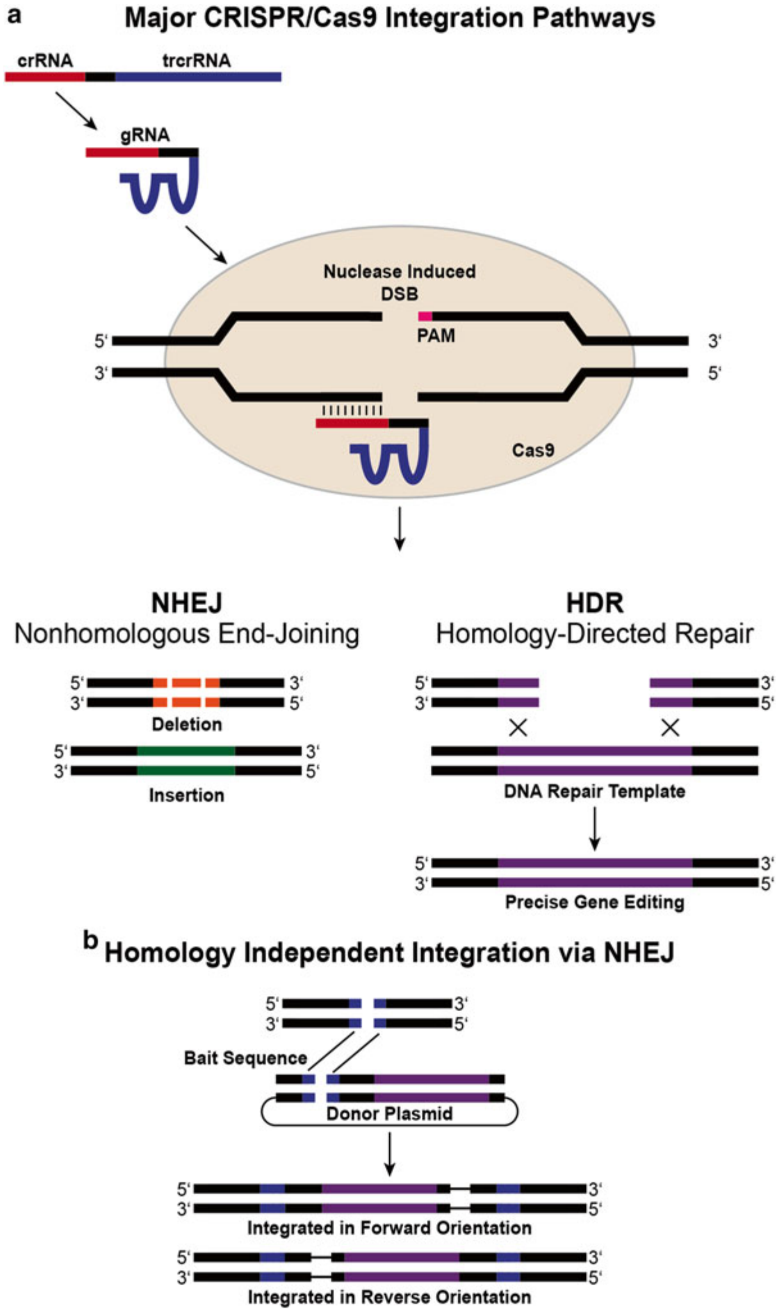

a
Major CRISPR/Cas9 Integration Pathways
crRNA
trcrRNA
gRNA
Nuclease Induced DSB
PAM
5'
3'
3'
5'
Cas9
NHEJ
Nonhomologous End-Joining
Deletion
Insertion
HDR
Homology-Directed Repair
DNA Repair Template
Precise Gene Editing
b
Homology Independent Integration via NHEJ
Bait Sequence
Donor Plasmid
Integrated in Forward Orientation
Integrated in Reverse Orientation

Fig. 15.3 CRISPR/Cas9 system. (**a**) To target a chromosomal locus a chimeric guide RNA, comprised of crRNA, which contains the homologous protospacer region and the tracrRNA, is used. The gRNA associates with Cas9 nuclease before it directs it to the desired chromosomal locus. Only loci that comprise a protospacer adjacent motive (PAM) will be successfully targeted. Cas9 induces DSB which can either be repaired through NHEJ or HDR. NHEJ is error prone and will induce insertions or deletions of variable length. HDR on the contrary will induce precise insertions to introduce genes or create specific mutations. (**b**) To utilize the NHEJ pathway for DNA insertion, a bait sequence on the donor vector adjacent to the integration DNA donor template that is also cut by the Cas9 nuclease, can be utilized (Adapted [58, 61])

donor template was suggested. While the bait sequence is used through the NHEJ pathway, the integration cassette can be correctly inserted. However, the orientation and integrity of the integration has to be verified [61] (Fig. 15.3c).

In summary, the CRISPR/Cas9 system allows specific targeting to a defined chromosomal location. While improvements still need to be made to reduce off-target effects and shift the balance towards HDR, the increased availability of whole genome sequences makes the CRISPR/cas9 technology actually the most promising and emerging strategy for genetic engineering of stable cell lines.

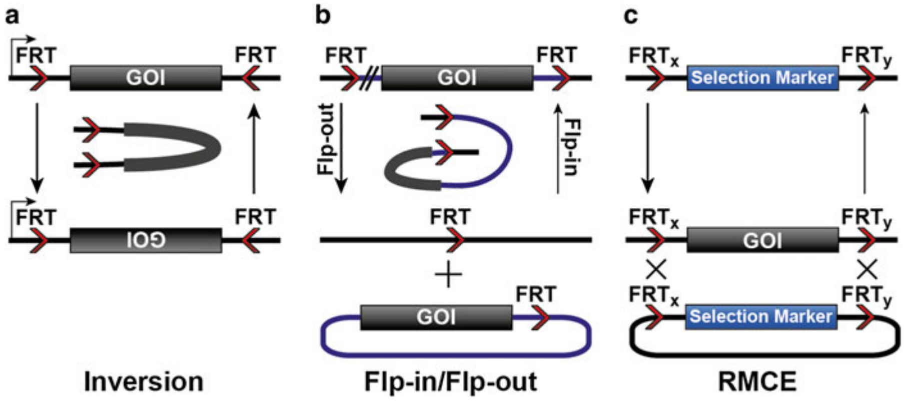


Fig. 15.4 Flp/*FRT* system. (**a**) Flp mediated inversion through anti parallel oriented homospecific FRT sites. (**b**) Flp-in and Flp-out reactions via unidirectional homospecific *FRT* sites. (**c**) Recombinase mediated cassette exchange (RMCE) introduces a gene of interest (GOI) between heterospecific *FRT* sites (Adapted from [66])

15.4.2 Site-Specific Recombination Systems

Site-specific recombination systems are commonly found in bacteria, bacteriophages and yeast. They catalyze DNA integration, excision or inversion through site specific recombination between specific DNA sequences (Fig. 15.4a, b) [62]. Since the early 1990s the tyrosine-type recombinases Flp and Cre are used for the targeted integration of transgenes into mammalian genomes [49, 50]. Flp variants with optimized thermo-stability (Flpe) [63] and codon usage (Flpo) [64] for use in mammalian systems became available as well as an array of mutant *FRT* sites [52, 65]. Therefore, the Flp/*FRT* system is currently our favored system for targeted integration [66].

In Flp/*FRT* based tag-and-target systems single *FRT* sites [49] or a homospecific set of *FRT* sites [67] are utilized for genomic tagging. However, both the Flp-in system [49] and the Flp-mediated DNA integration and rearrangement at prearranged genomic targets (FLIRT) system [67] will co-introduce prokaryotic vector elements which may induce epigenetic silencing. Recombinase mediated cassette exchange (RMCE), a tag-and-exchange strategy, does not co-introduce prokaryotic vector elements. Only the GOI and other vector elements such as promoters that are located within a set of heterospecific *FRT* sites on a donor vector will be introduced into a previously tagged chromosomal locus that comprises an exchange cassette with compatible *FRT* sites [52, 65] (Fig. 15.4c).

In our group the Flp/*FRT* system was utilized to generate stable CHO Lec3.2.8.1 cell lines [68, 69]. Initial efforts concentrated on the integration of one gene into one locus. The first strategy employed a tagging cassette that was comprised of homospecific *FRT*$_3$ sites flanking a fluorescent marker gene, an upstream promoter and a downstream GOI. Cells that successfully integrated the tagging cassette were isolated using fluorescent activated cell sorting (FACS). Upon isolation Flp mediated excision was used to remove the fluorescent marker gene (Fig. 15.5a). Thus the

Fig. 15.5 (continued) replaced the fluorescent marker. Furthermore, the inserted promoter and ATG recovered the downstream selection trap and allowed isolation of the producer cell line. (**c**) Binary RMCE: to obtain a binary RMCE master cell line a second exchange cassette comprising a tdTomato marker gene flanked by a different set of heterospecific *FRT* sites (*FRT*$_{13}$/*FRT*$_{14}$) and followed by a downstream puromycin selection trap (Δ*puro*) was randomly integrated into the genome of the eGFP positive RMCE master cell line. To isolate the binary RMCE master cell line high fluorescent tdTomato positive cells were isolated via FACS. Binary producer cell lines are obtained through the subsequential co-transfection of an exchange vector comprising the corresponding sets of *FRT* sites (either *FRT*$_3$/*FRT*$_{wt}$ or *FRT*$_{13}$/*FRT*$_{14}$) and the helper vector transcribing Flp recombinase

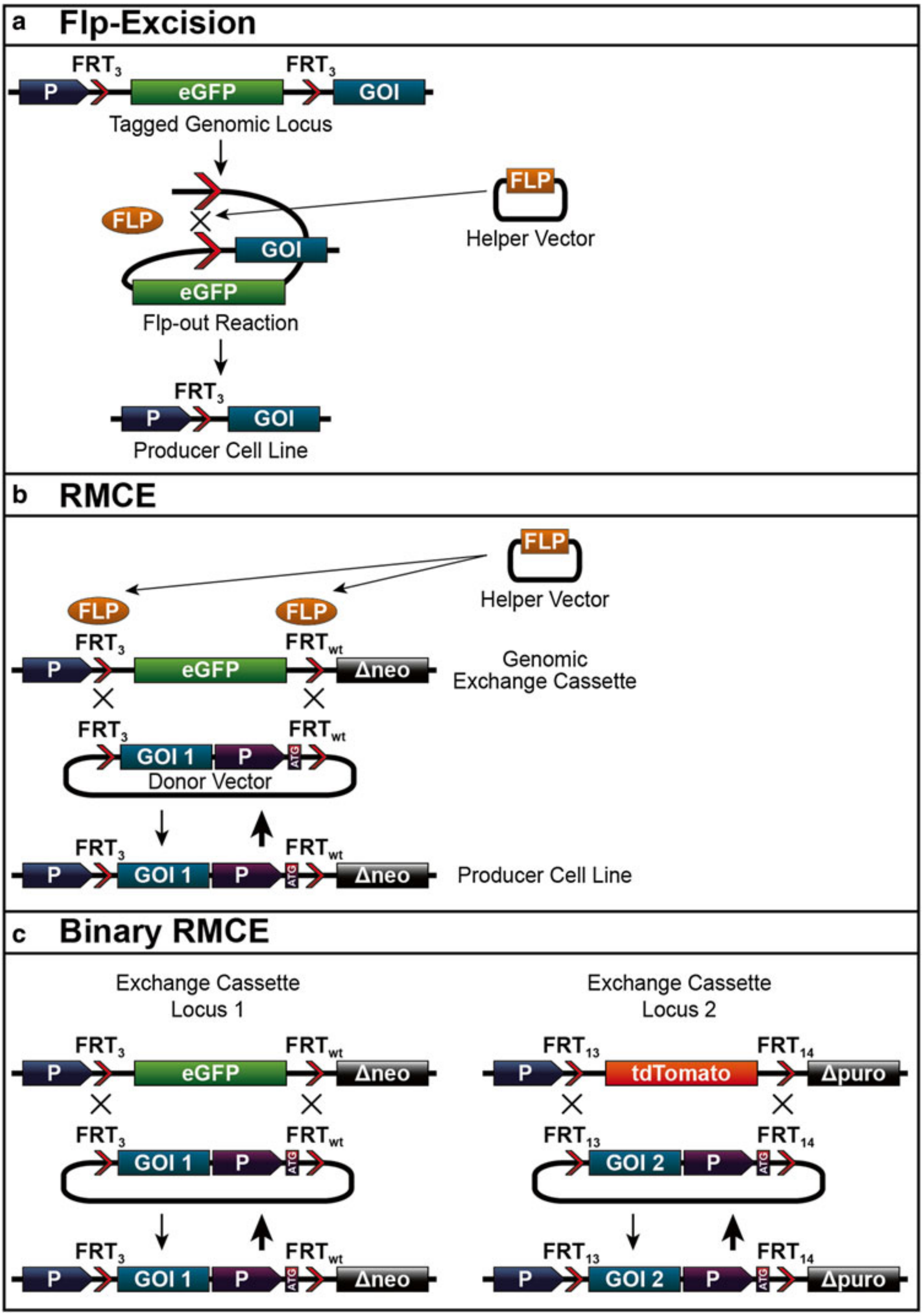


Fig. 15.5 Flp/*FRT* based systems. (**a**) Flp excision: a tagging cassette comprising homospecific FRT_3 sites flanking a fluorescent marker gene, an upstream promoter and a downstream GOI were stably integrated into the genome. The fluorescent marker gene was then removed via Flp mediated excision. Thus the GOI was placed under the control of the promoter leaving behind a single FRT_3 site. (**b**) RMCE: an exchange cassette containing an eGFP marker gene flanked by the heterospecific *FRT* sites: FRT_3/FRT_{wt} followed by a downstream neomycin selection trap (Δ*neo*) was integrated into the genome. RMCE master cell lines, which stably express eGFP were isolated by FACS and clonal selection. To generate producer cell lines, the master cell was co-transfected with an exchange vector comprising the GOI flanked by the corresponding set of *FRT* sites (FRT_3/FRT_{wt}) and a helper vector for production of Flp recombinase. Upon expression of Flp recombinase the cassette of the exchange vector

GOI was placed under the control of the promoter leaving behind a single *FRT*$_3$ site [68]. The second strategy was based on RMCE (Figs. 15.4c and 15.5b). An exchange cassette, comprising heterospecific *FRT* sites (*FRT*$_3$/*FRT*$_{wt}$) flanking a fluorescent marker gene with an upstream promoter and a downstream selection trap, was introduced into the host genome. Again high fluorescent positive cells were isolated with FACS to obtain stable master cell lines. Upon co-transfection of the master cell lines with a helper vector transcribing Flp recombinase and an exchange vector carrying the GOI, producer cell lines could be obtained. The expression cassette within the exchange vector comprises the GOI with a downstream promoter and an ATG start codon flanked by the same set of heterospecific *FRT* sites (*FRT*$_3$/*FRT*$_{wt}$) as in the tagged locus of the chromosome. The fluorescent marker gene is replaced upon successful integration by the expression cassette carrying the GOI. Additionally, the downstream promoter and the ATG start codon will restore the chromosomal resistant marker (selection trap). Applications of RMCE technology to generate stable cell lines in an array of mammalian cell lines have been reported by several research groups [69–73].

15.4.3 Binary RMCE

The previously described RMCE enables the targeted integration of a single GOI into a tagged genomic locus of a master cell line [52]. To enable the introduction of multiple transgenes in different chromosomal loci we established a binary RMCE system (Fig. 15.5c).

To enable the generation of multiRMCE master cell lines for the targeted integration of multiple transgenes into the genome, several requirements have to be fulfilled. A new set of tagging and exchange vectors with an alternative set of hetero-specific *FRT* sites, a second selection trap and an additional fluorescent selection marker have to be constructed [52, 65]. To circumvent the suppression or silencing of the expression of transgenes [13, 14], the expression cassettes should be located at different independent genomic loci. Binary RMCE master cell lines can be screened for integration loci with different expression characteristics. This enables the co-expression of transgenes at different relative ratios if required for efficient and balanced stoichiometric co-production of the protein partner.

To generate our binary RMCE master cell lines, we used the glycosylation mutant cell line CHO Lec3.2.8.1 [74] for the integration of two expression cassettes comprising different sets of heterospecific *FRT* sites (*FRT*$_3$/*FRT*$_{wt}$ and *FRT*$_{13}$/*FRT*$_{14}$). Fluorescence activated cell sorting (FACS) was shown to be the most suitable method for isolation of master cell lines without drug selection [75]. Each expression cassette was equipped with a different fluorescent maker gene (respectively eGFP or tdTomato) flanked by a set of heterospecific *FRT* sites to isolate the binary master cell line. To successfully isolate producer cell lines after RMCE both expression cassettes comprise a different downstream antibiotic selection trap, which lacks the promoter and the ATG start codon. The donor vector carrying only one pair of *FRT* sites supplies the desired GOI as well as a promoter and the ATG start codon for the selection trap. Upon integration of the *FRT*$_x$-GOI-Promoter-ATG-*FRT*$_y$ fragment into an exchange locus the master cell line will lose the corresponding fluorescent marker gene and gain resistance to the respective antibiotic selection marker (G418 or puromycin) (Fig. 15.5c).

For further improvement of the binary RMCE master cell lines, additional exchange loci with different heterospecific *FRT* sites could be integrated into the system to enable the integration of more than two genes at distinct chromosomal locations. Alternatively monocistronic expression cassettes can be designed which are able to subsequently introduce several GOIs into a given preselected locus. For example, a third heterospecific *FRT* site in front of the downstream *FRT* locus can be used to integrated additional expression cassettes at the high expression locus of the master cell line. Exchange vectors containing the following arrangement of *FRT* sites: *FRT*$_3$-Prom-GOI-Term-*FRT*$_x$/*FRT*$_{wt}$ will allow multiple rounds of integration and therefore further improve the multi-protein complex expression beyond the single and binary master cell lines by this new MultiRMCE strategy [76].

15.5 Summary

The elucidation of the structure and function of multi-subunit protein complexes is essential to understand their genuine position and task within the cellular network. Therefore, co-expression of all required subunits of the complex in mammalian systems is required if simpler bacterial expression hosts fail to efficiently over-express the desired target proteins. Transient as well as stable expression strategies in mammalian expression hosts can be used for recombinant protein production. To successfully assemble recombinant multi-subunit complexes all desired subunits optimally should be expressed within the same cell. Advances in fast generation of multi-gene plasmids, as by tandem recombineering, are particularly useful for transient protein expression but can also be used for the generation of stable mammalian cell lines. The isolation of stable high producer cell lines using random integration however is very time-consuming. Therefore targeted integration approaches that favor the integration into specific natural or previously tagged chromosomal loci are continuously improved. The use of engineered nucleases for integration of a transgene into a desired specific chromosomal locus is currently the most promising technology. Particularly the CRISPR/Cas9 technology will play an increasing role in generation of optimized producer cell lines. However, this technology has to be improved to reduce the number of off-target effects and to precisely modify and integrate transgenes into specific loci. Currently, targeted integration via previously tagged chromosomal loci through site-specific Cre or Flp recombination offers the most efficient way to generate stable cell lines with predictable expression properties.

The expression of multi protein complexes in mammalian cells is still challenging but the development of the described binary RMCE master cell lines and its further development as multiRMCE system, will have a major impact on the production of complex biologics for functional and structural studies in the near future.

Acknowledgments The authors received support by the Helmholtz Protein Sample Production Facility and by Instruct – Integrating Biology Infrastructure, part of the European Strategy Forum on Research Infrastructures (ESFRI). The authors also received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) for the ComplexINC Project under grant agreement N° 270089.

References

1. Robinson CV, Sali A, Baumeister W (2007) The molecular sociology of the cell. Nature 450(7172):973–982. doi:10.1038/nature06523
2. Mesa P, Deniaud A, Montoya G, Schaffitzel C (2013) Directly from the source: endogenous preparations of molecular machines. Curr Opin Struct Biol 23(3):319–325. doi:10.1016/j.sbi.2013.01.005
3. Schlatter S, Stansfield SH, Dinnis DM, Racher AJ, Birch JR, James DC (2005) On the optimal ratio of heavy to light chain genes for efficient recombinant antibody production by CHO cells. Biotechnol Prog 21(1):122–133. doi:10.1021/bp049780w
4. Li J, Zhang C, Jostock T, Dübel S (2007) Analysis of IgG heavy chain to light chain ratio with mutant Encephalomyocarditis virus internal ribosome entry site. Protein Eng Des Sel 20(10):491–496. doi:10.1093/protein/gzm038
5. Busso D, Peleg Y, Heidebrecht T, Romier C, Jacobovitch Y, Dantes A, Salim L, Troesch E, Schuetz A, Heinemann U, Folkers GE, Geerlof A, Wilmanns M, Polewacz A, Quedenau C, Bussow K, Adamson R, Blagova E, Walton J, Cartwright JL, Bird LE, Owens RJ, Berrow NS, Wilson KS, Sussman JL, Perrakis A, Celie PH (2011) Expression of protein complexes using multiple Escherichia coli protein co-expression systems: a benchmarking study. J Struct Biol 175(2):159–170. doi:10.1016/j.jsb.2011.03.004
6. Diebold ML, Fribourg S, Koch M, Metzger T, Romier C (2011) Deciphering correct strategies for multiprotein complex assembly by co-expression: application to complexes as large as the histone octamer. J Struct Biol 175(2):178–188. doi:10.1016/j.jsb.2011.02.001
7. Bieniossek C, Richmond TJ, Berger I (2008) MultiBac: multigene baculovirus-based eukaryotic protein complex production. Curr Protoc Protein Sci. doi:10.1002/0471140864.ps0520s51

8. Almo SC, Garforth SJ, Hillerich BS, Love JD, Seidel RD, Burley SK (2013) Protein production from the structural genomics perspective: achievements and future needs. Curr Opin Struct Biol 23(3):335–344. doi:10.1016/j.sbi.2013.02.014
9. Kerrigan JJ, Xie Q, Ames RS, Lu Q (2011) Production of protein complexes via co-expression. Protein Expr Purif 75(1):1–14. doi:10.1016/j.pep.2010.07.015
10. Tornøe J, Kusk P, Johansen TE, Jensen PR (2002) Generation of a synthetic mammalian promoter library by modification of sequences spacing transcription factor binding sites. Gene 297(1–2):21–32. doi:10.1016/S0378-1119(02)00878-8
11. Yang Y, Mariati, Ho SCL, Yap MGS (2009) Mutated polyadenylation signals for controlling expression levels of multiple genes in mammalian cells. Biotechnol Bioeng 102(4):1152–1160. doi:10.1002/bit.22152
12. Koh EY, Ho SC, Mariati, Song Z, Bi X, Bardor M, Yang Y (2013) An internal ribosome entry site (IRES) mutant library for tuning expression level of multiple genes in mammalian cells. PLoS One 8(12):e82100. doi:10.1371/journal.pone.0082100
13. Garrick D, Fiering S, Martin DI, Whitelaw E (1998) Repeat-induced gene silencing in mammals. Nat Genet 18(1):56–59. doi:10.1038/ng0198-56
14. Eszterhas SK, Bouhassira EE, Martin DIK, Fiering S (2002) Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. Mol Cell Biol 22(2):469–479. doi:10.1128/MCB.22.2.469-479.2002
15. Jäger V, Büssow K, Schirrmann T (2015) Transient recombinant potein expression in mammalian calls. In: Al-Rubeai M (ed) Animal cell culture, vol 9, 1st edn. Springer International Publishing, Switzerland, pp 27–64
16. Lebkowski JS, Clancy S, Calos MP (1985) Simian virus 40 replication in adenovirus-transformed human cells antagonizes gene expression. Nature 317(6033):169–171. doi:10.1038/317169a0
17. Young JM, Cheadle C, Foulke JS Jr, Drohan WN, Sarver N (1988) Utilization of an Epstein-Barr virus replicon as a eukaryotic expression vector. Gene 62(2):171–185. doi:10.1016/0378-1119(88)90556-2
18. Durocher Y, Perret S, Kamen A (2002) High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. Nucleic Acids Res 30(2):e9. doi:10.1093/nar/30.2.e9
19. Aricescu AR, Lu W, Jones EY (2006) A time- and cost-efficient system for high-level protein production in mammalian cells. Acta Crystallogr Sect D: Biol Crystallogr 62(Pt 10):1243–1250. doi:10.1107/s0907444906029799
20. Davies A, Greene A, Lullau E, Abbott WM (2005) Optimisation and evaluation of a high-throughput mammalian protein expression system. Protein Expr Purif 42(1):111–121. doi:10.1016/j.pep.2005.03.012
21. Geisse S, Henke M (2005) Large-scale transient transfection of mammalian cells: a newly emerging attractive option for recombinant protein production. J Struct Funct Genom 6(2-3):165–170. doi:10.1007/s10969-005-2826-4
22. Chaudhary S, Pak JE, Gruswitz F, Sharma V, Stroud RM (2012) Overexpressing human membrane proteins in stably transfected and clonal human embryonic kidney 293S cells. Nat Protoc 7(3):453–466. doi:10.1038/nprot.2011.453
23. Reeves PJ, Callewaert N, Contreras R, Khorana HG (2002) Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. Proc Natl Acad Sci U S A 99(21):13419–13424. doi:10.1073/pnas.212519299
24. Shukla S, Schwartz C, Kapoor K, Kouanda A, Ambudkar SV (2012) Use of baculovirus BacMam vectors for expression of ABC drug transporters in mammalian cells. Drug Metab Dispos 40(2):304–312. doi:10.1124/dmd.111.042721
25. Dukkipati A, Park HH, Waghray D, Fischer S, Garcia KC (2008) BacMam system for high-level expression of recombinant soluble and membrane glycoproteins for structural studies. Protein Expr Purif 62(2):160–170. doi:10.1016/j.pep.2008.08.004
26. Ramos L, Kopec LA, Sweitzer SM, Fornwald JA, Zhao H, McAllister P, McNulty DE, Trill JJ, Kane JF (2002) Rapid expression of recombinant proteins in modified CHO cells using the baculovirus system. Cytotechnology 38(1-3):37–41. doi:10.1023/A:1021189628274
27. Kost TA, Condreay JP, Ames RS, Rees S, Romanos MA (2007) Implementation of BacMam virus gene delivery technology in a drug discovery setting. Drug Discov Today 12(9-10):396–403. doi:10.1016/j.drudis.2007.02.017
28. Kost TA, Condreay JP, Jarvis DL (2005) Baculovirus as versatile vectors for protein expression in insect and mammalian cells. Nat Biotechnol 23(5):567–575. doi:10.1038/nbt1095
29. Wurm FM (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. Nat Biotechnol 22(11):1393–1398. doi:10.1038/nbt1026
30. Oberbek A, Matasci M, Hacker DL, Wurm FM (2011) Generation of stable, high-producing cho cell lines by lentiviral vector-mediated gene transfer in serum-free suspension culture. Biotechnol Bioeng 108(3):600–610. doi:10.1002/bit.22968
31. Appleby SL, Irani Y, Mortimer LA, Brereton HM, Klebe S, Keane MC, Cowan PJ, Williams KA (2013) Co-expression of a scFv antibody fragment and a reporter protein using lentiviral shuttle plasmid containing a self-processing furin-2A sequence. J Immunol Methods 397(1–2):61–65. doi:10.1016/j.jim.2013.08.012

32. Mufarrege EF, Antuna S, Etcheverrigaray M, Kratje R, Prieto C (2014) Development of lentiviral vectors for transient and stable protein overexpression in mammalian cells. A new strategy for recombinant human FVIII (rhFVIII) production. Protein Expr Purif 95:50–56. doi:10.1016/j.pep.2013.11.005
33. Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. Nat Biotechnol 22(12):1583–1587. doi:10.1038/nbt1036
34. Fitzgerald DJ, Berger P, Schaffitzel C, Yamada K, Richmond TJ, Berger I (2006) Protein complex expression by using multigene baculoviral vectors. Nat Methods 3(12):1021–1032. doi:10.1038/nmeth983
35. Bieniossek C, Nie Y, Frey D, Olieric N, Schaffitzel C, Collinson I, Romier C, Berger P, Richmond TJ, Steinmetz MO, Berger I (2009) Automated unrestricted multigene recombineering for multiprotein complex production. Nat Methods 6(6):447–450. doi:10.1038/nmeth.1326
36. Kriz A, Schmid K, Baumgartner N, Ziegler U, Berger I, Ballmer-Hofer K, Berger P (2010) A plasmid-based multigene expression system for mammalian cells. Nat Commun 1(8):120. doi:10.1038/ncomms1120
37. Vijayachandran LS, Viola C, Garzoni F, Trowitzsch S, Bieniossek C, Chaillet M, Schaffitzel C, Busso D, Romier C, Poterszman A, Richmond TJ, Berger I (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J Struct Biol 175(2):198–208. doi:10.1016/j.jsb.2011.03.007
38. Craig A, Berger I (2011) ACEMBL expression system series *Multi*Mam multi-protein expression in mammalian cells. Version 1.0
39. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T (2005) Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell 122(3):473–483. doi:10.1016/j.cell.2005.07.013
40. Wu SC-Y, Meir Y-JJ, Coates CJ, Handler AM, Pelczar P, Moisyadi S, Kaminski JM (2006) PiggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. Proc Natl Acad Sci 103(41):15008–15013. doi:10.1073/pnas.0606979103
41. Ivics Z, Izsvak Z (2010) The expanding universe of transposon technologies for gene and cell engineering. Mob DNA 1(1):25. doi:10.1186/1759-8753-1-25
42. Wilson MH, Coates CJ, George AL (2006) PiggyBac transposon-mediated gene transfer in human cells. Mol Ther 15(1):139–145. doi:10.1038/sj.mt.6300028
43. Matasci M, Baldi L, Hacker DL, Wurm FM (2011) The piggyBac transposon enhances the frequency of CHO stable cell line generation and yields recombinant lines with superior productivity and stability. Biotechnol Bioeng 108(9):2141–2150. doi:10.1002/bit.23167
44. Li Z, Michael IP, Zhou D, Nagy A, Rini JM (2013) Simple piggyBac transposon-based mammalian cell expression system for inducible protein production. Proc Natl Acad Sci U S A 110(13):5004–5009. doi:10.1073/pnas.1218620110
45. Balasubramanian S, Matasci M, Kadlecova Z, Baldi L, Hacker DL, Wurm FM (2015) Rapid recombinant protein production from piggyBac transposon-mediated stable CHO cell pools. J Biotechnol. doi:10.1016/j.jbiotec.2015.03.001
46. Kaji K, Norrby K, Paca A, Mileikovsky M, Mohseni P, Woltjen K (2009) Virus free induction of pluripotency and subsequent excision of reprogramming factors. Nature 458(7239):771–775. doi:10.1038/nature07864
47. Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, Hämäläinen R, Cowling R, Wang W, Liu P, Gertsenstein M, Kaji K, Sung H-K, Nagy A (2009) piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. Nature 458(7239):766–770. doi:10.1038/nature07863
48. Kahlig KM, Saridey SK, Kaja A, Daniels MA, George AL Jr, Wilson MH (2010) Multiplexed transposon-mediated stable gene transfer in human cells. Proc Natl Acad Sci U S A 107(4):1343–1348. doi:10.1073/pnas.0910383107
49. O'Gorman S, Fox D, Wahl G (1991) Recombinase-mediated gene activation and site-specific integration in mammalian cells. Science 251(4999):1351–1355. doi:10.1126/science.1900642
50. Fukushige S, Sauer B (1992) Genomic targeting with a positive-selection lox integration vector allows highly reproducible gene expression in mammalian cells. Proc Natl Acad Sci 89(17):7905–7909. doi:10.1073/pnas.89.17.7905
51. Huang LC, Wood EA, Cox MM (1997) Convenient and reversible site-specific targeting of exogenous DNA into a bacterial chromosome by use of the FLP recombinase: the FLIRT system. J Bacteriol 179(19):6076–6083
52. Schlake T, Bode J (1994) Use of mutant FLP TRecognition Target (FRT) sites for the exchange of expression cassettes at defined chromosomal loci. Biochemistry 33(43):12746–12751. doi:10.1021/bi00209a003
53. Gaj T, Gersbach CA, Barbas CF 3rd (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31(7):397–405. doi:10.1016/j.tibtech.2013.04.004
54. Moehle EA, Rock JM, Lee YL, Jouvenot Y, DeKelver RC, Gregory PD, Urnov FD, Holmes MC (2007) Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. Proc Natl Acad Sci U S A 104(9):3055–3060. doi:10.1073/pnas.0611478104
55. Urnov FD, Miller JC, Lee Y-L, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature 435(7042):646–651. doi:10.1038/nature03556

56. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339(6121):819–823. doi:10.1126/science.1231143
57. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. Science 339(6121):823–826. doi:10.1126/science.1232033
58. Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 32(4):347–355. doi:10.1038/nbt.2842
59. Hsu Patrick D, Lander Eric S, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. Cell 157(6):1262–1278. doi:10.1016/j.cell.2014.05.010
60. Lee JS, Kallehauge TB, Pedersen LE, Kildegaard HF (2015) Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. Sci Rep 5:8572. doi:10.1038/srep08572
61. Auer TO, Del Bene F (2014) CRISPR/Cas9 and TALEN-mediated knock-in approaches in zebrafish. Methods 69(2):142–150. doi:10.1016/j.ymeth.2014.03.027
62. Hirano N, Muroi T, Takahashi H, Haruki M (2011) Site-specific recombinases as tools for heterologous gene integration. Appl Microbiol Biotechnol 92(2):227–239. doi:10.1007/s00253-011-3519-5
63. Buchholz F, Angrand P-O, Stewart AF (1998) Improved properties of FLP recombinase evolved by cycling mutagenesis. Nat Biotechnol 16(7):657–662. doi:10.1038/nbt0798-657
64. Raymond CS, Soriano P (2007) High-efficiency FLP and ΦC31 site-specific recombination in mammalian cells. PLoS ONE 2(1):e162. doi:10.1371/journal.pone.0000162
65. Turan S, Kuehle J, Schambach A, Baum C, Bode J (2010) Multiplexing RMCE: versatile extensions of the Flp-recombinase-mediated cassette-exchange technology. J Mol Biol 402(1):52–69. doi:10.1016/j.jmb.2010.07.015
66. Turan S, Galla M, Ernst E, Qiao J, Voelkel C, Schiedlmeier B, Zehe C, Bode J (2011) Recombinase-Mediated Cassette Exchange (RMCE): traditional concepts and current challenges. J Mol Biol 407(2):193–221. doi:10.1016/j.jmb.2011.01.004
67. Huang Y, Li Y, Wang Y, Gu X, Wang Y, Shen B (2007) An efficient and targeted gene integration system for high-level antibody expression. J Immunol Methods. doi:10.1016/j.jim.2007.01.022
68. Wilke S, Krausze J, Gossen M, Groebe L, Jäger V, Gherardi E, van den Heuvel J, Büssow K (2010) Glycoprotein production for structure analysis with stable, glycosylation mutant CHO cell lines established by fluorescence-activated cell sorting. Protein Sci 19(6):1264–1271. doi:10.1002/pro.390
69. Wilke S, Groebe L, Maffenbeier V, Jäger V, Gossen M et al (2011) Streamlining homogeneous glycoprotein production for biophysical and structural applications by targeted cell line development. PLoS ONE 6(12):e27829, doi: 27810.21371/journal.pone.0027829
70. Seibler J, Schubeler D, Fiering S, Groudine M, Bode J (1998) DNA cassette exchange in ES cells mediated by Flp recombinase: an efficient strategy for repeated modification of tagged loci by marker-free constructs. Biochemistry 37(18):6229–6234. doi:10.1021/bi980288t
71. Nehlsen K, Schucht R, da Gama-Norton L, Kromer W, Baer A, Cayli A, Hauser H, Wirth D (2009) Recombinant protein expression by targeting preselected chromosomal loci. BMC Biotechnol 9(1):100. doi:10.1186/1472-6750-9-100
72. Mayrhofer P, Kratzer B, Sommeregger W, Steinfellner W, Reinhart D, Mader A, Turan S, Qiao J, Bode J, Kunert R (2014) Accurate comparison of antibody expression levels by reproducible transgene targeting in engineered recombination-competent CHO cells. Appl Microbiol Biotechnol. doi:10.1007/s00253-014-6011-1
73. Meyer S, Lorenz C, Baser B, Wördehoff M, Jäger V, van den Heuvel J (2013) Multi-host expression system for recombinant production of challenging proteins. PLoS ONE 8(7):e68674. doi:10.1371/journal.pone.0068674
74. Stanley P (1989) Chinese hamster ovary cell mutants with multiple glycosylation defects for production of glycoproteins with minimal carbohydrate heterogeneity. Mol Cell Biol 9(2):377–383. doi:10.1128/mcb.9.2.377
75. Liu W, Xiong Y, Gossen M (2006) Stability and homogeneity of transgene expression in isogenic cells. J Mol Med 84(1):57–64. doi:10.1007/s00109-005-0711-z
76. Turan S, Zehe C, Kuehle J, Qiao J, Bode J (2013) Recombinase-mediated cassette exchange (RMCE) – a rapidly-expanding toolbox for targeted genomic modifications. Gene 515(1):1–27. doi:10.1016/j.gene.2012.11.016

Part V

Plant Expression

Microalgae as Solar-Powered Protein Factories

16

Franziska Hempel and Uwe G. Maier

Abstract

Microalgae have an enormous ecological relevance as they contribute significantly to global carbon fixation. But also for biotechnology microalgae became increasingly interesting during the last decades as many algae provide valuable natural products. Especially the high lipid content of some species currently attracts much attention in the biodiesel industry. A further application that emerged some years ago is the use of microalgae as expression platform for recombinant proteins. Several projects on the production of therapeutics, vaccines and feed supplements demonstrated the great potential of using microalgae as novel low-cost expression platform. This review provides an overview on the prospects and advantages of microalgal protein expression systems and gives an outlook on potential future applications.

Keywords

Bioreactor • Microalgae • Protein expression platform • Recombinant proteins

16.1 Introduction

In 1982 human insulin was the first protein that was produced in a microbial system and approved for pharmaceutical use. Today recombinant proteins are indispensible in daily life, as they became essential instruments in many industrial sectors like food, fuel, textile and pharma industry. Especially the medical sector is a fast growing market and complex eukaryotic proteins like monoclonal antibodies, hormones and growth factors are needed in high quantity and dominated the biotech field for the last years reaching US

F. Hempel
LOEWE Center for Synthetic Microbiology
(SYNMIKRO),
Hans-Meerwein-Strasse, Marburg 35043, Germany

U.G. Maier (✉)
LOEWE Center for Synthetic Microbiology
(SYNMIKRO),
Hans-Meerwein-Strasse, Marburg 35043, Germany

Laboratory for Cell Biology, Philipps-Universität
Marburg,
Karl-von-Frisch Strasse 8, Marburg 35043, Germany
e-mail: maier@biologie.uni-marburg.de

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_16

sales of more than 40 billion dollar in 2011 [1]. Unfortunately, the production costs for most of these proteins are still very high limiting a broad therapeutic use, hence present expression systems need to be improved and novel production platforms should be explored.

Bacteria were the first expression systems for recombinant proteins established already in the 1970s and today still represent the basic workhorse in white biotechnology. They are first choice for production of most industrial enzymes and also 30 % of pharmaceutical proteins are still produced in bacteria as growth rates and expression levels are very high and overall production costs are relatively low [2]. For production of most complex eukaryotic proteins, however, bacterial systems are not feasible as proteins lack eukaryotic post-translational modifications critical for folding, stability and biological activity e.g. disulfide bond formation, phosphorylation and glycosylation being most important modifications [3]. Furthermore, many eukaryotic proteins accumulate in bacteria as insoluble aggregates and need costly downstream processing for purification and refolding. Some of these problems can be overcome by the use of yeasts like *S. cerevisiae*. These fungi are able to perform eukaryotic post-translational modifications and are as cost-effective as bacteria exhibiting high growth rates, high productivity and easy scalability [4]. However, a major problem concerning production of pharmaceutical proteins in yeast is linked to N-linked glycosylation, which differs from mammalian systems and hinders expression of correctly modified complex human therapeutics. Inefficient secretion and proteolysis are further critical issues for high scale expression of complex therapeutics like monoclonal antibodies [5]. Nevertheless, advances in metabolic engineering e.g. towards humanized glycosylation patterns might make yeast more interesting for pharmaceutical protein production in future [6–9]. Other eukaryotic expression systems like insect cells exist but 50 % of licensed pharmaceutical proteins are currently produced in mammalian cells, i.e. hamster cell lines (CHO), human cell lines or hybridoma cell lines in case of monoclonal antibodies [2, 10]. The major advantage of these expression systems is that recombinant proteins produced in mammalian cell lines exhibit correct post-translational modifications needed for therapeutic applications. On the other side, however, mammalian systems are very limited in scale up options and the production process is very cost-intense due to expensive media and complex cultivation processes representing serious bottlenecks for high scale production pipelines [11]. The contamination with human pathogens is a further critical issue necessitating thorough checks on biosafety [12].

To overcome high production costs and reduce the risk of pathogenic contaminations the idea of using plants as expression platform for recombinant proteins became very popular in the 1990s and is often referred to as *molecular farming* [13–16]. Plant-based production is fueled by sunlight, thus the production process itself should be very cheap and agricultural cultivation is well established. Ideally, plants might be used as production platform for edible vaccines making such therapeutics available for large parts of the population and especially in developing countries where they are needed most [17, 18]. In the last 25 years a lot of effort was put into that field of research and trials with engineered plant systems producing humanized glycosylation patterns look promising [19, 20]. However, major hurdles for using whole plants as expression system are low production levels and costly purification processes for products with no oral application. Furthermore, ethical concerns as well as the risk of contaminating food crops make the cultivation of transgenic plants a highly controversial issue [21]. Plant cell culture based systems appear more promising as higher production levels and protein secretion can be achieved and contained reactors allow production with good manufacturing practice (GMP) [22, 23]. In 2012 human glucocerebrosidase produced in a carrot cell culture system was the first plant made pharmaceutical approved by the Food and Drug Administration (FDA) [1]. However, cultivation costs in plant cell cultures are still significant.

16.2 Algae as Bioreactor for Recombinant Protein Production

Algae are solar-powered like plants and especially unicellular microalgae that possess high photosynthesis rates and yield biomass much faster than plants are interesting for diverse biotechnological applications. Many microalgae species provide valuable natural compounds such as vitamins, pigments, proteins and lipids and have been used in cosmetic, food and veterinary industry for many years [24, 25]. Microalgae attract currently much attention in the biofuel sector as many species possess high lipid content and might provide a sustainable and cheap source for biodiesel in future [26–29]. Still underestimated though, is the idea of using microalgae as expression platform for recombinant proteins. Microalgae can be cultivated with low costs needing basically water and sunlight and combine rapid growth rates, easy handling and high scale up capacity with the advantages of eukaryotic expression systems [30–33]. The genome sequence of different microalgae species became available within the last years and basic genetic tools like stable transfection of the nucleus and chloroplast genome and inducible promoter systems were established to express recombinant proteins within different cellular compartments or target proteins for secretion into the culture medium. Compared to complex systems like plants or mammalian cells, microalgae are very robust and easily accessible for genetic manipulation making rapid high-throughput analysis possible. Many microalgae are used as food source and are regarded as safe as they contain no harmful components and are no host for human pathogens. Hence, the expression of therapeutic proteins and also oral application of whole cell extracts represent a promising option. Within the last years different therapeutic proteins like monoclonal antibodies, immunotoxins, subunit vaccines and feed supplements have been expressed in microalgae [31]. Most of these studies focused so far on the green alga *Chlamydomonas reinhardtii*, which is the model alga in basic research and was the first microalga to be sequenced and accessible for genetic engineering. At present *C. reinhardtii* is mainly used for recombinant protein expression in the chloroplast, but also other algal systems like *Dunaliella* and *Chlorella* species and the diatom *Phaeodactylum tricornutum* are now explored and especially *P. tricornutum* reveals great potential in using alga for expressing recombinant proteins from the nucleus genome. This review highlights recent progress in using microalgae as expression system for recombinant proteins and gives an overview on general concepts and practical considerations.

16.3 Pro- and Eukaryotic Expression Traits Within One Cell

In contrast to most other expression systems microalgae provide the opportunity to express recombinant proteins either from the nucleus genome in a eukaryotic environment or in an “advanced” prokaryotic milieu within the chloroplast. Most research so far focused on expression in the chloroplast as in the model alga *C. reinhardtii* higher expression levels were observed for the chloroplast than for the nucleus genome ranging from 0.1 to 5 % of total soluble protein and even 21 % in one study (Table 16.1). Another interesting feature for protein expression within the chloroplast is the fact that this originally prokaryotic organelle, unlike bacteria, harbors an advanced set of chaperons and enzymes to form disulfide bonds [73, 74]. Complete IgG antibodies and special variants like dimeric single chain antibodies can be produced in the chloroplast of *C. reinhardtii* and are fully assembled and able to bind to the target antigens [36, 39]. Recently, it was shown that the chloroplast is also interesting for the production of immunotoxins, chimeric proteins that consist of an antigen-binding domain fused to a eukaryotic toxin like PE40 or gelonin [37, 38]. As these components are toxic for eukaryotes that kind of therapeutics normally have to be produced in bacterial systems with the drawback that protein complexity is very limited. In the chloroplast, however, complex divalent

Table 16.1 Overview of algal produced therapeutics, feed supplements and other recombinant proteins (Modified and updated from Rasala et al. [31])

Recombinant protein		Algae species	Expression site	Expression level	Genome, promoter	Comments	Reference
Antibodies							
Human IgG αHBsAg	Antibody against Hepatitis B Virus surface protein	*P. tricornutum*	Intracellular (ER)	~9 % TSP, 22 mg/g dw	Nucleus, NR	IgG antibodies are fully assembled and recognize the target antigen; high levels of recombinant protein were obtained without large scale screening or genetic engineering of wild type cells	[34]
			Secreted	2.5 mg/L		Fully assembled and functional IgG antibodies are secreted efficiently into the culture medium; rarely other proteins were secreted hence the antibody in the media was already relatively pure	[35]
Human IgG αPA83	Antibody against Anthrax	*C. reinhardtii*	Intracellular (chloroplast)	100 μg/g dw	Chloroplast, *psbA*	Assembled and functional IgG antibodies accumulate in the chloroplast demonstrating that this prokaryotic organelle is able to fold also complex eukaryotic molecules	[36]
Immunotoxin αCD22-PE40	scFv antibodies against CD22 coupled to PE40	*C. reinhardtii*	Intracellular (chloroplast)	0.2–0.4 % TSP	Chloroplast, *psbA*	Mono- and dimeric single chain immunotoxins were able to bind and kill B cell lymphoma cells in vitro; anti-tumor activity was observed in mice	[37]
Immunotoxins αCD22-Gelonin	scFv antibodies against CD22 coupled to Gelonin	*C. reinhardtii*	Intracellular (chloroplast)	0.1–0.7 % TSP	Chloroplast, *psbA*	Mono- and dimeric single chain immunotoxins were able to bind and prevent proliferation of B-cell lymphomas; the inhibitory effect of dimeric immunotoxins was 15 to 25-fold higher compared to the monomeric form	[38]
lsc αHSV glycoprotein D	Large single-chain antibody against Herpes simplex virus glycoprotein D	*C. reinhardtii*	Intracellular (chloroplast)	n.s.	Chloroplast, *atpA*, *rbcL*	First report on antibody production in a microalga; Large single chain antibodies were expressed in the chloroplast and were proven to assemble as dimer that binds to target antigen	[39]

Nanobodies αBoNT/A	Variable domains of camelid heavy chain antibodies against botulinum neurotoxin A	*C. reinhardtii*	Intracellular (chloroplast)	5 % TSP	Chloroplast, *psbA*	Nanobodies bound to BoNT/A in ELISA and were capable of protecting rat neurons from BoNT/A inactivation; nanobodies stayed intact within the stomach and small intestine of mice after oral delivery of transgenic algae	[40]
Vaccines							
D2-CTB	D2 fibronectin binding domain of S. aureus fused to ß-subunit of Cholera toxin	*C. reinhardtii*	Intracellular (chloroplast)	0.7 % TSP 1.6 mg/g dw	Chloroplast, *rbcL*	Oral vaccinated mice show mucosal and systemic immune response and are protected from lethal *S. aureus* infections; lyophilized vaccine is active for at least 1.5 years at room temperature	[41]
AMA1- and MSP1-GBSS	*P. berghei* Apical Major Antigen 1 and Major Surface Protein 1 fused to the algal granule bound starch synthase	*C. reinhardtii*	Intracellular (chloroplast)	0.2–1 μg/ mg starch	Nucleus, *hsp70A-rbcS2*	Malaria vaccines were targeted to starch granules; oral as well as intraperitoneally vaccination led to reduced parasitemia in mice; MSP1 elicited IgG antibodies inhibiting intra-erythrocytic asexual development of different Plasmodium species *in vitro*	[42]
Pfs25, Pfs28	*P. falciparum* surface proteins 25 and 28	*C. reinhardtii*	Intracellular (chloroplast)	0.5 %, 0.2 % TSP	Chloroplast, *psbA*	Both antigens are soluble and correctly folded; algal produced Pfs25 is shown to elicit antibodies with transmission blocking activity in mice	[43]
Psf25-CTB	*P. falciparum* surface protein 25 fused to the β-subunit of cholera toxin	*C. reinhardtii*	Intracellular (chloroplast)	0.09 % TSP	Chloroplast, *psbA*	Oral vaccination of mice elicited CTB and Pfs25 specific IgA antibodies but no transmission blocking IgG antibodies; CTB-Pfs25 is stable in freeze-dried algae at 4–22 °C for at least 6 month	[44]
Pfs48/45 C-term	*P. falciparum* surface protein 48/45, C-terminal domain	*C. reinhardtii*	Intracellular (chloroplast)	n.s.	Chloroplast, *psbA*, *psbD*	C-terminal part of Pfs 48/45 is recognized by transmission blocking antibodies	[45]
Bovine LFB-dsRed	Bovine lactoferricin, anti-microbial peptide	*N. oculata*	n.s.	n.s.	Nucleus, *hsp70A-rbcs2* (*C. reinhardtii*)	Oral vaccination of medaka fish resulted in significant survival rate after *Vibrio parahaemo-lyticus* infection	[46]

(continued)

Table 16.1 (continued)

Recombinant protein		Algae species	Expression site	Expression level	Genome, promoter	Comments	Reference
VP28 (WSSV)	Vaccine against White Spot Syndrome Virus	*D. salina*	Intracellular (n.s.)	0.0003 % TSP	Nucleus, Ω TMV	Oral vaccination of crayfish elicited significant survival rate after WSSV infection	[47]
HBsAg	Hepatitis B Virus surface antigen	*P. tricornutum*	Intracellular (ER)	0.7 % TSP	Nucleus, NR	Algal produced HBsAg is recognized by αHBsAg antibodies in inhibitory ELISA	[34]
		D. salina	Intracellular (n.s.)	1.6–3.1 ng/mg TSP	Nucleus, ubil-Ω	First report on recombinant protein expression in *D. salina*; no assays on functionality provided	[48]
E7 HPV-16	E7 oncogene of Human Papilloma Virus 16	*C. reinhardtii*	Intracellular (chloroplast)	0.12 % TSP	Chloroplast, *psbD*	Subcutaneous vaccination of mice elicits specific IgG response, T cell proliferation and tumor protection	[49]
VP1 (FMDV)-CTB	Vaccine against Foot-and-Mouth-Disease Virus fused to ß-subunit of Cholera toxin	*C. reinhardtii*	Intracellular (chloroplast)	3 % TSP	Chloroplast *chlL*	Algal produced protein was shown to bind to GM1 ganglioside receptors; no in vivo studies for oral vaccination were performed	[50]
VP28 (WSSV)	Vaccine against White Spot Syndrome Virus	*C. reinhardtii*	Intracellular (chloroplast)	0.2–21 % TCP	Chloroplast, *psbA*, *atpA*	Highest expression level for microalgal chloroplast transfection reported so far; large scale screens revealed highly variable expression levels with same transfection parameters	[51]
E2 (CSFV)	Vaccine against Classical Swine Fever Virus	*C. reinhardtii*	Intracellular (chloroplast)	1.5–2 % TSP	Chloroplast, *rbcL*	Subcutaneous immunization with algal extract elicits serum specific antibodies in mice	[52]
Angiotensin II-HBcAg	Vaccine against hypertension	*C. reinhardtii*	Intracellular (n.s.)	0.05 % TSP	Nucleus, CaMV35S	Angiotensin II was expressed in fusion with Hepatitis B Virus nucleocapsid antigen	[53]
Other therapeutic proteins							
hGH	Human growth hormone	*C. vulgaris*, *C. sorokiniana*	Secreted	200–600 ng/mL	Nucleus, CaMV35S, *rbcS2*	One of the first reports on expression of a therapeutic protein in micro-algae; no stable transformants	[54]

M-SAA	Bovine mammary-associated serum amyloid, gut active therapeutic	*C. reinhardtii*	Intracellular (chloroplast)	5 % TSP	Chloroplast, *psbD*, *psbA*	Purified algal produced protein stimulated mucin expression in cell culture demonstrating potential as edible gut active therapeutic acting in prophylaxis of bacterial and viral infections	[55]
Human Epo	Erythropoietin, therapeutic for anemia treatment	*C. reinhardtii*	Secreted	~100 µg/L	Nucleus, *hsp70A-rbcS2*	Epo is secreted into the culture medium when fused with an endogenous signal peptide; intronic sequences enhance the expression of recombinant proteins	[56]
Epo, interferon-ß, proinsulin, VEGF, HMGB1, 10FN3, 14FN3	Therapeutics for diverse treatments	*C. reinhardtii*	Intracellular (chloroplast)	Up to 3 % TSP	Chloroplast, *psbA*, *atpA*	All algal produced proteins were soluble and showed biological activity; expression level was enhanced under psbA promoter	[57]
Rabbit NP-1	Rabbit neutrophil peptide 1, anti-microbial peptide	*C. ellipsoidea*	Intracellular (n.s.)	n.s.	Nucleus, ubiquitin-Ω	Anti-microbial activity of algal extract against different bacteria and fungi was confirmed	[58]
		C. ellipsoidea	Intracellular (n.s.)	11.4 mg/L culture	Nucleus, ubiquitin-Ω	Purified NP-1 protein exhibited strong anti-microbial activity against *E. coli*	[59]
Human TRAIL	TNF-related apoptosis-inducing ligand induces apoptosis in various tumor cells	*C. reinhardtii*	Intracellular (chloroplast)	0.43–0.67 % TSP	Chloroplast, *atpA*	TRAIL is expressed as soluble protein	[60]
hGAD65	Autoantigen human glutamic acid decarboxylase 65; marker for diabetes I diagnostics	*C. reinhardtii*	Intracellular (chloroplast)	0.25–0.3 % TSP	Chloroplast, *rbcL*	Purified algal produced hGAD65 induced proliferation of spleen cells from NOD mice and showed immunoreactivity to diabetic sera	[61]
SKTI	Soybean kunitz trypsin inhibitor, therapeutic protein with anti-viral and anti-cancer activity	*D. salina*	Intracellular (n.s.)	0.68 % TSP	Nucleus, 35S	Stable expression; no assays on functionality	[62]

(continued)

Table 16.1 (continued)

Recombinant protein		Algae species	Expression site	Expression level	Genome, promoter	Comments	Reference
apcA + apcB	Allophycocyanin, potential therapeutic against S-180 carcinoma	*C. reinhardtii*	Intracellular (chloroplast)	2–3 % TSP	Chloroplast, *atpA*	Both subunits were expressed; assembly of the protein complex not verified	[63]
Feed supplements							
fGH	Flounder growth hormone, feed additive for aquaculture	*C. ellipsoidea*	Intracellular, (n.s.)	400 μg/L culture	Nucleus, CaMV35S	Flounder fry fed on transformed *Chlorella* revealed 25 % growth increase after 30 days	[64]
ypGH	Yellowfin porgy growth hormone, feed additive for aquaculture	*N. oculata*	Intracellular (n.s.)	0.27–0.41 μg/mL culture	Nucleus, *hsp70A-rbcS2* (*C. reinhardtii*)	Red tilapia fry fed on transformed *N. oculata* showed 212 % weight increase and 71 % length increase after 4 weeks	[65]
Phytase AppA	Feed additive to increase phytate phosphorus utilization	*C. reinhardtii*	Intracellular (chloroplast)	n.s.	Chloroplast, *atpA*	Orally delivered algae extract reduced fecal phytate excretion in broiler chicks	[66]
Endo-β-1,4-xylanase	Feed additive for hemicellulose breakdown	*C. reinhardtii*	Intracellular (cytosol) Secreted	0.25 % TSP n.s.	Nucleus, AR4	Expression of xyn1 as fusion protein with the selection marker led to 100-fold increase of xyn1 levels; self-cleaving peptide allowed production and secretion of autonomous xyn1 functional in activity assays	[67]
Xylanase, α-Galactohy-drolase, phytase	Div. feed additives	*D. tertiolecta*	Intracellular (chloroplast)	n.s.	Chloroplast, *psbD*	Algal produced enzymes were functional in activity assays and expression was compared to *C. reinhardtii* expression; first description of *D. tertiolecta* chloroplast transformation protocol	[68]

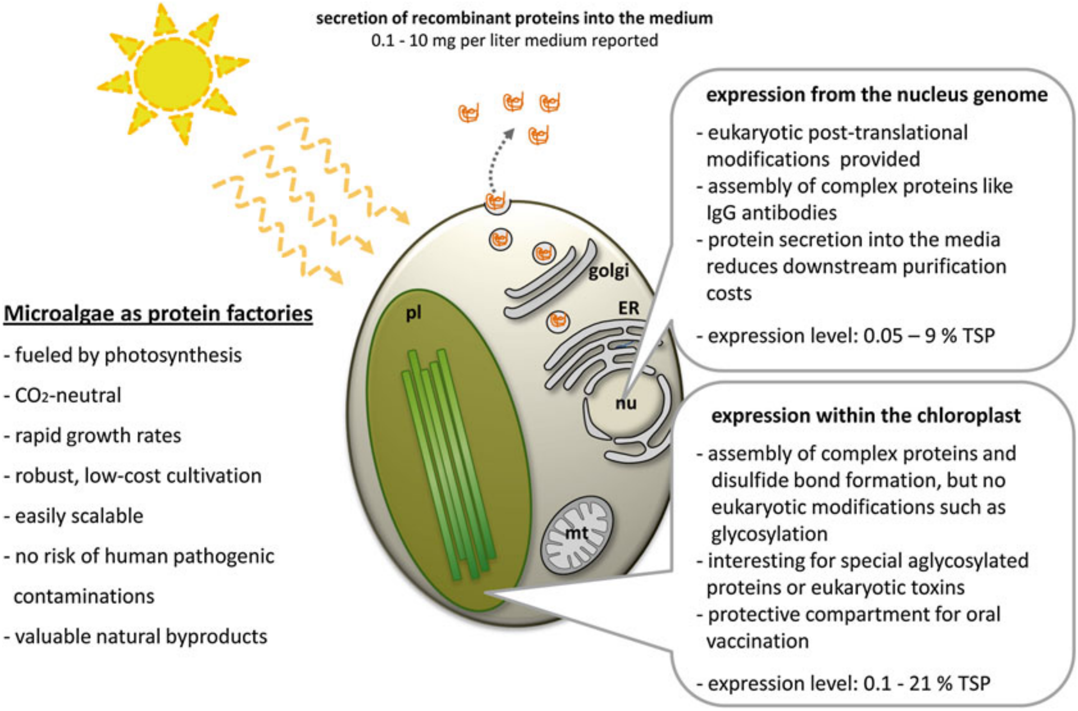
hu Sep15	Selenoprotein, nutritional supplement	*C. reinhardtii*	Intracellular (n.s.)	n.s.	Nucleus, *hsp70-rbcS2*	Cell viability was not affected	[69]
Bioremediation and environmental control							
Metallothionein-like protein (*F. rubra*)	Metal binding	*C. reinhardtii*	Intracellular (chloroplast)	n.s.	Chloroplast, *atpA*	Transgenic algae showed higher cadmium binding capacity and tolerance; IC_{50} of Cd^{2+} was 55.43 % enhanced compared to wild type strain	[70]
TMOF	Trypsin-modulating oostatic factor, insecticide against mosquito larvae	*Chlorella* sp.	Intracellular	?	?	Feeding to mosquito larvae caused larval mortality	[71]
hMT-2	Human metallothionein-2, protection against UV-radiation	*C. reinhardtii*	Intracellular (chloroplast)	n.s.	Chloroplast, *psbA*	Expression of hMT-2 conferred enhanced resistance to UV-B exposure	[72]

CTB cholera toxin subunit B, *dw* dry weight, *ER* endoplasmic reticulum, *NR* nitrate reductase, *n.s.* not specified, *TCP* total chloroplast protein, *TSP* total soluble protein immunotoxins can be produced showing enhanced cytotoxicity compared to the monovalent form [37]. Beside antibodies, further complex structured eukaryotic proteins like *Plasmodium falciparum* peptides interesting as malaria transmission-blocking vaccine were shown to be expressed as soluble and correctly folded proteins within the chloroplast [43–45]. Most proteins of *P. falciparum* are not glycosylated and expression in classical eukaryotic expression systems is problematic [75]. As chloroplasts harbor no glycosylation machinery, proteins expressed within the chloroplast remain aglycosylated, which represents a further advantage in this special case preventing allergic reactions to foreign glycoprofiles. Feeding mice with freeze-dried algae expressing *P. falciparum* surface protein 25 elicited specific IgA and IgG antibodies demonstrating that algae are like plants highly interesting in terms of oral vaccination [44]. Protein storage within the chloroplast might come along with enhanced protein protection as in another study a chloroplast expressed *Staphylococcus aureus* protein was shown to be potent in lyophilized algae for at least 20 month at room temperature and protected against proteolysis within a stomach-like pepsin environment [41]. In summary, the algal chloroplast is of special interest for the expression of complex aglycosylated or toxic proteins that cannot be produced in bacterial systems. Oral delivery might represent an ideal application form for different subunit vaccines as well as for some feed supplements [31, 76] (Fig. 16.1).

Protein expression from the nucleus genome provides the advantage of eukaryotic post-translational modifications, which are important for conformation, stability and activity of most eukaryotic proteins. Nevertheless expression from algal nucleus genomes was so far mostly disregarded as most studies in *C. reinhardtii* showed rather low expression levels ranging from 0.05 to 0.25 % of total soluble protein (Table 16.1). Research especially from the last years, however, revealed that this topic deserves more attention and that it is worth to test other algal species as well. In 2011 a human IgG antibody against the Hepatitis B Virus Surface protein was expressed in the diatom *P. tricornutum* and accumulated to ~9 % of total soluble protein—corresponding to about 22 mg antibody per 1 g dry weight [34]. Interestingly, further analyses revealed that the deletion of an initially used ER-retention signal even led to secretion of the fully assembled and functional IgG antibodies into the culture medium [35]. As *P. tricornutum* does not seem to secrete many proteins by natural means, the antibodies were remarkably pure and accumulated in this assay without further engineering to about 2.5 mg/L medium [35]. The secretion of proteins into the culture medium is of course a great benefit simplifying downstream processes and reducing cost-intense purification steps incredibly. Recent work demonstrated that also engineered *C. reinhardtii* cells can secrete high amounts of recombinant protein into the medium as shown exemplarily for the reporter protein gLuc (*Gaussia princeps* luciferase) yielding approximately 10 mg per liter culture medium in case of an engineered cell wall deficient strain [77]. Altogether, research especially from the last years reveals the potential of microalgae as solar-driven expression system capable to produce complex eukaryotic proteins that can be secreted efficiently into the culture media. The chloroplast as an advanced prokaryotic compartment represents a further expression site within the algal cell interesting for special products like complex aglycosylated proteins or eukaryotic toxins (Fig. 16.1).

16.4 Genetic Engineering for Enhanced Protein Expression in Microalgae

Transformation of the green alga *C. reinhardtii* was established about 25 years ago [78, 79] and today many other algae, *e.g.*, the green algae species *Dunaliella* and *Chlorella* as well as Heterokontophytes like *P. tricornutum*, *Thalassiosira pseudonana* and *Nannochloropsis* species can be routinely transformed. Microparticle bombardment is widely used for introducing new genes into the nucleus or chloroplast genome but also electroporation can be applied for most species



secretion of recombinant proteins into the medium
0.1 - 10 mg per liter medium reported

Microalgae as protein factories

- fueled by photosynthesis
- CO_2-neutral
- rapid growth rates
- robust, low-cost cultivation
- easily scalable
- no risk of human pathogenic contaminations
- valuable natural byproducts

expression from the nucleus genome

- eukaryotic post-translational modifications provided
- assembly of complex proteins like IgG antibodies
- protein secretion into the media reduces downstream purification costs
- expression level: 0.05 – 9 % TSP

expression within the chloroplast

- assembly of complex proteins and disulfide bond formation, but no eukaryotic modifications such as glycosylation
- interesting for special aglycosylated proteins or eukaryotic toxins
- protective compartment for oral vaccination
- expression level: 0.1 - 21 % TSP

Fig. 16.1 Microalgae as solar powered expression system for recombinant proteins. Microalgae possess rapid growth rates, are very robust as well as easily scalable and might provide low-cost production of different therapeutic proteins as well as feed supplements in future. Proteins can be expressed either from the nucleus genome or within the chloroplast. Both options provide advantages depending on the application and the protein of interest. Most studies concentrated on the green alga *C. reinhardtii* so far but also other systems like *P. tricornutum*, *Chlorella* and *Dunaliella* species start to get explored and might provide some advantages. *ER* endoplasmatic reticulum, *mt* mitochondrion, *nu* nucleus, *pl* plastid

(Table 16.2) representing a more economical approach for large-scale assays. For detailed reviews on algal transformation techniques and selectable marker genes see Gong et al. [100], Leon-Banares et al. [101], and Potvin and Zhang [102]. In general, chloroplast transformation techniques take advantage of homologous recombination to integrate gene constructs into a specific genomic context [68, 78, 84]. Integration into the nucleus genome occurs in case of most transformation techniques random via non-homologous end joining. Only for *N. oceanica* it was shown so far that homologous recombination works also for the nucleus genome [86] representing a great benefit as the integration context can influence gene expression considerably. Targeted gene knock out is a further great advantage as it offers many engineering options possible in other algae like *C. reinhardtii* and *P. tricornutum* only via RNA interference at the moment [103, 104]. Very recently, however, also in these organisms some progress has been made for targeted insertion and gene knock out regarding the nuclear genome as engineered zinc-finger nucleases have been shown to allow specific DNA insertion in *C. reinhardtii* [105] and also for *P. tricornutum* genomic modifications via meganucleases and TALE nucleases have been reported [106, 107].

Like in other expression systems also for microalgae the adaption of DNA-sequences to the host specific codon-usage is beneficial and turned out to be very important at least for *C. reinhardtii* (with a GC-content of 61 %) in order to obtain detectable protein levels [108–110]. For the diatom *P. tricornutum*, which possesses a typical eukaryotic GC-content of 48 % this might be less critical. Foreign genes can also be expressed without codon optimization and even bacterial enzymes can be expressed from the nucleus genome of *P. tricornutum* as shown in a study on bioplastic production [111].

The choice of the promoter is a further critical feature for efficient protein expression and regarding expression from the chloroplast genome 5′UTRs of *psbA*, *psbD*, *atpA* and *rbcL* are most frequently used (Table 16.2). The *psbA* promoter is one of the most efficient in *C. reinhardtii*, at least when the endogenous *psbA* gene is deleted [51, 55, 112], and expression levels of 0.1–21 % TSP were observed (Table 16.1). Photosynthesis can be restored in these strains by reintroducing an attenuated *psbA* gene in a different gene context [55]—however production rates decreased and strains were still unviable for commercial scale. Recently, *psbA* complementation was optimized, though, allowing high phototrophic growth rates while maintaining high production levels [113]. For recombinant protein expression from the nucleus genome of *C. reinhardtii* the 5′UTRs of *hsp70A*, *psaD*, *rbcS2* or the fusion *hsp70A*-*rbcS2* have been used in most assays (Table 16.1), however expression levels are very low, which might be a problem of gene silencing. Best expression levels were observed in a genetic screen of a mutant library accumulating recombinant protein to 0.2 % of TSP [114]. For a more detailed review on chloroplast and nuclear promoter studies in *C. reinhardtii* see Specht et al. [115]. The diatom *P. tricornutum* came into focus as expression system for recombinant proteins only very recently and therefore less data is available, but much higher nuclear expression levels than in *C. reinhardtii* were observed in initial tests when using the inducible promoter of the endogenous nitrate reductase (8.7 % of TSP and 0.7 %) [34, 35]. This promoter was established previously in the diatom *T. pseudonana* and can be tightly controlled via ammonia/nitrate in the culture medium [116]. The light inducible promoters of *fcpA* and *fcpB* of *P. tricornutum* are frequently used in basic research but no quantifications on expression levels are available so far.

Further strategies that have shown to enhance nuclear expression of recombinant proteins in microalgae include the insertion of introns from native genes [56, 117] and the expression as transcriptional fusion to an antibiotic resistance gene [67, 109]. Both strategies were applied in *C. reinhardtii* and might help to counteract transgene silencing. For the industrial enzyme xylanase it was exemplarily shown that expression levels could increase to 100-fold when expressing this protein in fusion with the coding region for the selection marker. The insertion of a viral self-cleavable peptide guaranteed a discrete protein as final product, which could also be secreted when including an endogenous signal peptide [67].

Altogether, the molecular toolbox for microalgae was more and more extended within the last years and not only the model alga *C. reinhardtii* but also other species like *P. tricornutum* and *N. oceanica* start to get explored revealing beneficial traits like higher expression levels from the nucleus genome or targeted gene insertion, respectively. In future, additional engineering might help to increase expression levels and provide modifications such as humanized glycosylation profiles—a critical quality attribute with the basics just being about to get investigated in microalgae [118–120]. In future, also other algal species like *Chlorella* might become interesting as expression platform as *Chlorella* possesses very rapid growth rates and provides valuable natural compounds interesting for food, cosmetic and biodiesel industry (Table 16.2). Compared to other microalgae, however, molecular tools for the expression of recombinant proteins in *Chlorella* are still very limited and studies on recombinant protein expression are still in a very early stage.

16.5 Expression of Protein Complexes

Multi-protein complexes are essential for many cellular processes and for their structural and biochemical analyses as well as for therapeutic applications large-scale production is necessary. However, the heterologous expression of protein

Table 16.2 Selected microalgal species interesting for the production of recombinant proteins

Microalga	Natural habitat	Relevant products	Genome sequence	Transformation techniques	Molecular tools
Phaeodactylum tricornutum (Heterokontophyta)	Sea water	Omega-3-fatty acids (nutrition); Lipids (biofuel)	+ [80]	Nucleus: biolistic [81]; electroporation [82, 83]; chloroplast: electroporation [84]	Inducible promoters, gene knock-down via antisense RNA. Gene knockout via TALEN, secretion of recombinant proteins confirmed
Nannocloropsis oceanica (Heterokontophyta)	Sea water	Polysaccharides, proteins, vitamins (nutrition); Lipids (biofuel)	+ [85]	Nucleus: electroporation [86]; electroporation [65]; electroporation [87]	Inducible promoters, gene knockout via homologous recombination
N. oculata			In progress		
N. gaditana			+ [87]		
Thalassiosira pseudonana (Heterokontophyta)	Sea water	Silica (nanotechnology)	+ [88]	Nucleus: biolistic [89]; biolistic [79]; glass beads [91]; electroporation [92]; *Agrobacterium* [93]; chloroplasts: biolistic [78]	Inducible promoters, gene knock-down via antisense RNA
Chlamydomonas reinhardtii (Chlorophyta)	Fresh water		+ [90]		Inducible promoters, gene knock-down via antisense RNA, Gene knockout via ZFN, secretion of recombinant proteins confirmed
Chlorella vulgaris (Chlorophyta)	Fresh water	Polysaccharides, proteins, vitamins (cosmetics, nutrition); Lipids (biofuel)	In progress *C. variabilis* [94]	Nucleus: electroporation [95]	Limited tools available, in progress, secretion of recombinant proteins confirmed
C. ellipsoidea			–	Nucleus: electroporation [58]; biolistic [96]	
C. sorokiniana			–		
Dunaliella salina (Chlorophyta)	Sea water	β-carotene and other carotenoids (cosmetics, nutrition); Lipids (biofuel)	In progress	Nucleus: electroporation [48], biolistic [97], glass beads [98]	Limited tools available, in progress
D. tertiolecta			–	Nucleus: electroporation [99]; chloroplast: biolistic [68]	

complexes consisting of multiple protein subunits still represents a great challenge. Historically, protein subunits were initially expressed separately, purified and reconstituted *in vitro*, but of course this is problematic in many cases as proteins form aggregates, have to be refolded and additional factors that might be needed for the assembly process are not present. Today co-expression and complex formation within the cell is favored and different complexes have been successfully expressed in bacteria, yeast, insect cells, plants as well as some mammalian cell lines. Different strategies are applied like the use of multiple expression vectors, plasmids with multiple cloning sites or polycistronic units in case of bacteria, as well as the expression of fusion proteins (see Kerrigan et al. [121] for a review). Data on the expression of protein complexes in microalgae is so far very limited but as mentioned previously the efficient expression of completely assembled IgG complexes consisting of two heavy and two light chains was shown to be feasible in *P. tricornutum* [34, 35]. The complex is assembled within the ER and can be secreted into the culture medium. Also in the chloroplast of *C. reinhardtii* completely assembled IgG complexes can be produced. The expression of other protein complexes has yet not been tested in microalgae; however, the molecular tools for the co-expression of multiple protein subunits are basically available. In *P. tricornutum* multiple plasmids can be co-transformed [122]; this technique was applied for example to introduce three bacterial enzymes for production of the bioplastic PHB [111]. As different resistance markers are available for many algae sequential transfections can be performed as well. Also plasmids with multiple cloning sites like the plasmid pPha-DUAL-[2xNR], which contains two multiple cloning sites, both under the control of a nitrate-inducible promoter, are available and have been used for the expression of IgG antibodies in *P. tricornutum* [34]. In *C. reinhardtii* it was shown very recently that also fusion proteins separated by viral self-cleaving sequences can be expressed from the nucleus genome leading to separate gene products [67, 123]. Hence, basic tools for the expression of multiple protein subunits in microalgae are available and it will be highly interesting to start expression assays of multi-subunit complexes in future.

16.6 Algal Produced Therapeutics, Feed Supplements and Other Proteins

Within the last 15 years a broad spectra of recombinant proteins has been produced in different microalgae ranging from therapeutic proteins like antibodies, vaccines, hormones to feed supplements and industrial relevant enzymes (Table 16.1). The following sections provide an overview on different studies in the field.

16.6.1 Antibodies

Therapeutic antibodies currently represent the best-selling class of biologics with US sales reaching 20.3 billion dollar in 2011 [1]. Antibody production is based on mammalian cell culture, which is very expensive and therefore alternative expression systems are highly desirable. In 2003, a large single-chain antibody against Herpes simplex virus glycoprotein D was the first antibody to be expressed in an algal system [39]. This antibody was produced in the chloroplast of *C. reinhardtii* and was proven to assemble as dimer that binds to its target antigen. Further studies on antibody expression include the production of a complete human IgG antibody against anthrax in the chloroplast of *C. reinhardtii* [36] and the expression of a human IgG antibody against the Hepatitis B Virus surface protein. In contrast to previous studies, the latter was expressed from the nucleus genome in the diatom *P. tricornutum* and was produced very efficiently with 9 % of total soluble protein [34]. The deletion of a retention signal led to efficient secretion of the fully assembled and functional antibodies into the culture medium [35]. As rarely other proteins were detected in the media, the antibody was relatively pure without further treatment. In 2013, the production of mono and dimeric single chain immunotoxins was shown to be feasible in the chloroplast of *C. reinhardtii*. These algal produced antibody variants were able to bind and kill B cell lymphoma cells *in vitro* and showed anti tumor activity in mice [37, 38]. Very recently, also camelid antitoxins were expressed in the chloroplast of *C. reinhardtii*. The algal produced nanobodies against botulinum neurotoxin A showed *in vitro* activity in toxin protection assays and remained intact in the gastrointestinal tract of mice fed with antitoxin-producing microalgae [40]. Altogether, these studies demonstrate the great potential of using microalgae as expression system for antibodies and further engineering concerning productivity and glycopatterns might make microalgae a low-cost production platform with little risk for pathogenic contaminations in future.

16.6.2 Vaccines

Protein based vaccines represent a further important class of therapeutics and microalgae might be of interest especially for the production of oral subunit vaccines. Many algal species are used as nutritional supply in food industry, hence complete cell extracts could be administered directly thereby bypassing costly purification steps and facilitating production as well as needle-free application [76]. Different reports on the expression of oral vaccines in microalgae have been published within the last years. In 2010, the D2 fibronectin-binding domain of *S. aureus* was expressed in fusion with the adjuvants CTB (cholera toxin B subunit) in the chloroplast of *C. reinhardtii*. Orally vaccinated mice showed mucosal as well as systemic immune response resulting in protection from lethal *S. aureus* infections [41]. The vaccine was stable within lyophilized algae for at least 20 month at room temperature. Furthermore, stability assays demonstrated that the protein was protected against proteolysis within a stomach-like pepsin environment representing a critical point for absorption within the intestine. The study demonstrates that the algal chloroplast could represent an ideal compartment for the expression of oral vaccines resulting in enhanced antigen stability and protection. In another study *Plasmodium berghei* antigens were targeted to chloroplast starch granules in *C. reinhardtii* [42]. Oral vaccination of mice led to reduced parasitemia and specific IgG antibodies could be detected inhibiting intra-erythrocytic asexual development of different Plasmodium species *in vitro*. A further study on the production of a malaria transmission blocking vaccine in the chloroplast of *C. reinhardtii*, the *Plasmodium falciparum* surface protein 25 (Psf25-CTB), demonstrated that the protein is correctly folded and elicits transmission blocking antibodies in mice when injected intraperitoneally [43]. Orally vaccinated mice produced specific mucosal IgA antibodies but no systemic IgG antibody production was observed [44]. Of course, not every vaccine is suitable for oral administration to generate a systemic immune response and also the adjuvants used in this study might not be ideal for stimulating IgG production. But even though oral vaccination is still very limited due to the complexity of the mucosal system, ongoing research will certainly help to overcome present challenges with microalgae representing promising edible, low-cost expression systems [75]. Also in veterinary medicine and especially in aqua culture using microalgae as vaccination vehicle might be of great interest. One of the first reports on the production of oral therapeutics in microalgae is the expression of the anti-microbial peptide bovine lactoferricin in *Nannochloropsis oculata* [46]. Feeding experiments with medaka fish revealed a survival rate of 85 % after *Vibrio parahaemolyticus* infection. In another study the protein VP28 of the White Spot Syndrome Virus was expressed in the green alga *Dunaliella salina*. Even though expression levels were only very low, oral vaccination of crayfish resulted in significant survival rates after white spot syndrome virus (WSSV) infection [47].

Other vaccines that were produced in microalgae but not tested for oral application are the hepatitis B virus surface antigen (HBsAg) produced in the chloroplast of *D. salina* as well as in the endoplasmic reticulum of *P. tricornutum* [34, 48]. Additionally, different *Plasmodium* antigens interesting for malaria control [43, 45] and proteins of different viruses, *i.e.*, human papilloma virus 16 [49], white spot syndrome virus [51], foot and mouth disease virus [50] and classical swine fever virus [52], were expressed within the chloroplast of *C. reinhardtii* with expression levels of 0.12–21 % of total soluble protein (Table 16.1).

16.6.3 Other Therapeutic Proteins

Human growth hormone (hGH) was one of the first therapeutic proteins that were tested for expression in an algal system. In 1999, when molecular engineering of microalgae was still in the very beginning, Hawkins and colleagues expressed hGH in *Chlorella vulgaris* as well as in *Chlorella sorokiniana*, and showed that the hormone is expressed and secreted into the culture medium, even though transfection was only transient in these initial studies [54]. The bovine mammary-associated serum amyloid protein (M-SAA) is an anti-microbial protein that is found in the colostrum (first milk) of mammals and prevents bacterial infections by stimulating mucin synthesis in the small intestines. M-SAA was shown to be expressed efficiently in the chloroplast of *C. reinhardtii* as bioactive molecule stimulating mucin production in epithelial cell culture [55]. The study demonstrates the great potential of using microalgae for the production of edible gut active therapeutics and only recently, the US company Triton Health and Nutrition started the microalgal production of M-SAA. Further therapeutics that were expressed in algal systems include human erythropoietin, which was produced in the chloroplast of *C. reinhardtii* [57] as well as shown to be secreted when expressed from the nucleus genome [56]. Furthermore, diverse therapeutics such as interferon-β, proinsulin [57] and a marker for diabetes I diagnostics [61] were expressed in the chloroplast of *C. reinhardtii*. All proteins were shown to be biologically active and accumulated to up to 3 % of total soluble protein. The production of anti-microbial peptides was also tested in *D. salina* and *C. ellipsoidea* [58, 59, 62], which might become interesting as an expression system when more molecular tools become available in future.

16.6.4 Animal Feed Supplements

Many algae are used as feed additives since many species are rich in valuable natural compounds like long-chain polyunsaturated fatty acids (PUFAs), carotenoids, vitamins and high quality protein and carbohydrate. The opportunity to combine nutritional supply and direct delivery of recombinant feed additives like growth hormones or dietary enzymes for fiber break down makes microalgae an interesting low-cost expression system in that field. In 2002 the flounder growth hormone (fGH) was expressed in *C. ellipsoidea* and it was shown that flounder fry fed on transformed algae exhibit a 25 % growth increase after 1 month [64]. Promising results were also obtained when feeding tilapia fry with transgenic *N. oculata* cells expressing yellowfin porgy growth hormone (ypGH) [65]. After 4 weeks, a 212 % increase in weight and 71 % increase in length were observed. In another study, a bacterial phytase was expressed in *C. reinhardtii* to facilitate phytate digestion in monogastric animals. Feeding experiments on broiler chicks revealed a reduced excretion of fecal phytate [66]. The production of other phytases and further enzymes used as dietary supplements like α-galactohydrolases and a xylanase was also shown to be feasible in *C. reinhardtii* and *D. tertiolecta* [26].

16.6.5 Bioremediation and Environmental Control

Algae are used in wastewater treatment to provide oxygen supply for bacterial biodegradation and remove inorganic nitrogen and phosphorus [124]. In addition, the removal of heavy metals like cadmium, nickel and zinc was shown for some species [125, 126] and might be enhanced by genetic engineering. In two studies the expression of metallothioneins was assayed in *C. reinhardtii* demonstrating that transgenic cells exhibit higher cadmium binding capacity and can grow to higher densities at toxic cadmium concentrations [70, 127]. Another study presents an algal-based approach on mosquito control. An insecticide against mosquito larvae acting on trypsin biosynthesis in the mosquito gut was produced in *Chlorella* sp. [71]. Feeding of transgenic algae to mosquito larvae caused larval mortality.

Altogether, a broad repertoire of recombinant proteins like many different therapeutic proteins, feed supplements for animal welfare and proteins interesting for environmental control have been expressed in microalgae (Table 16.1) demonstrating the great potential of microalgal expression systems.

16.7 Conclusions and Perspectives

The demand for recombinant protein therapeutics and industrial enzymes is enormous nowadays and the market for biologics is constantly growing. In future, it will be essential to improve existing expression systems but also to establish novel low-cost production platforms to guarantee affordable products. Microalgae are powered by sunlight, possess rapid growth rates and are genetically well accessible. In recent years significant progress has been made in recombinant protein expression in microalgae and many different projects on the production of therapeutics, vaccines and feed supplements demonstrate the great potential of using microalgae as novel low-cost expression platform. Microalgae are of special interest for the production of complex eukaryotic proteins that currently have to be produced in mammalian cell lines involving high production costs and the risk of human pathogenic contaminations. Especially the production of IgG antibodies that were shown to be secreted into the algae culture medium could offer an attractive option in future. The expression of oral vaccines and feed supplements represent a further promising approach. As many microalgae are edible and harbor valuable vitamins, proteins and fatty acids the complete cell extract could be administered directly saving expensive purification costs. Furthermore, the algal chloroplast represents an interesting expression platform especially for the production of complex aglycosylated proteins or eukaryotic toxins coupled to complex proteins like bivalent immunotoxins that are difficult to produce in other expression systems.

Concerning commercial applications, the production of recombinant proteins in microalgae is still at an early stage. Comparable to the establishment of previous expression systems it will be necessary to enhance productivity and secretion efficiency and establish glycoengineering approaches to provide therapeutics with humanized glycoprofiles in future. In addition to *C. reinhardtii* other microalgal strains like *P. tricornutum*, *N. oceanica* and *Chlorella* should be included for detailed expression studies as it was started in some projects only recently. Higher expression levels, exclusive molecular tools and higher growth rates could be some of the advantages. As microalgae biotechnology became very popular within the last years, a lot of effort has also been put into the development of algae photobioreactors. Promising solutions for large-scale cultivation can be provided by now [128] representing a further important aspect concerning technology transfer to industrial scale in future. Triton Health and Nutrition (USA) and Algenics (France) belong to the first companies that started to use microalgae as expression platform for recombinant proteins. But certainly other companies will follow soon considering the great potential of using microalgae as solar fueled, low-cost expression system.

Acknowledgments We are grateful to Dr. Stefan Zauner (Marburg) for comments on the manuscript. This work was supported by the LOEWE program of the state of Hesse, Germany.

References

1. Aggarwal SR (2012) What's fueling the biotech engine-2011 to 2012. Nat Biotechnol 30(12):1191–1197
2. Ferrer-Miralles N, Domingo-Espin J, Corchero JL, Vazquez E, Villaverde A (2009) Microbial factories for recombinant pharmaceuticals. Microb Cell Factories 8:17

3. Walsh G, Jefferis R (2006) Post-translational modifications in the context of therapeutic proteins. Nat Biotechnol 24(10):1241–1252
4. Martinez JL, Liu L, Petranovic D, Nielsen J (2012) Pharmaceutical protein production by yeast: towards production of human blood proteins by microbial fermentation. Curr Opin Biotechnol 23(6):965–971
5. Frenzel A, Hust M, Schirrmann T (2013) Expression of recombinant antibodies. Front Immunol 4:217
6. Amano K, Chiba Y, Kasahara Y, Kato Y, Kaneko MK, Kuno A, Ito H, Kobayashi K, Hirabayashi J, Jigami Y, Narimatsu H (2008) Engineering of mucin-type human glycoproteins in yeast cells. Proc Natl Acad Sci U S A 105(9):3232–3237
7. De Pourcq K, De Schutter K, Callewaert N (2010) Engineering of glycosylation in yeast and other fungi: current state and perspectives. Appl Microbiol Biotechnol 87(5):1617–1631
8. Wildt S, Gerngross TU (2005) The humanization of N-glycosylation pathways in yeast. Nat Rev Microbiol 3(2):119–128
9. Ye J, Ly J, Watts K, Hsu A, Walker A, McLaughlin K, Berdichevsky M, Prinz B, Sean Kersey D, d'Anjou M, Pollard D, Potgieter T (2011) Optimization of a glycoengineered Pichia pastoris cultivation process for commercial antibody production. Biotechnol Prog 27(6):1744–1750
10. el Redwan RM (2007) Cumulative updating of approved biopharmaceuticals. Hum Antibodies 16(3–4):137–158
11. Dietmair S, Nielsen LK, Timmins LE (2012) Mammalian cells as biopharmaceutical production hosts in the age of omics. Biotechnol J 7(1):75–89
12. Pauwels K, Herman P, Van Vaerenbergh B, Dai Do thi C, Berghmans L, Waeterloos G, Van Bockstaele D, Dorsch-Häsler K, Sneyers M (2007) Animal cell cultures: risk assessment and biosafety recommendations. Appl Biosaf 12(1):26–38
13. Daniell H, Singh ND, Mason H, Streatfield SJ (2009) Plant-made vaccine antigens and biopharmaceuticals. Trends Plant Sci 14(12):669–679
14. Fischer R, Stoger E, Schillberg S, Christou P, Twyman RM (2004) Plant-based production of biopharmaceuticals. Curr Opin Plant Biol 7(2):152–158
15. Franken E, Teuschel U, Hain R (1997) Recombinant proteins from transgenic plants. Curr Opin Biotechnol 8(4):411–416
16. Ma JK, Drake PM, Christou P (2003) The production of recombinant pharmaceutical proteins in plants. Nat Rev Genet 4(10):794–805
17. Fooks AR (2000) Development of oral vaccines for human use. Curr Opin Mol Ther 2(1):80–86
18. Giddings G, Allison G, Brooks D, Carter A (2000) Transgenic plants as factories for biopharmaceuticals. Nat Biotechnol 18(11):1151–1155
19. Gomord V, Fitchette AC, Menu-Bouaouiche L, Saint-Jore-Dupas C, Plasson C, Michaud D, Faye L (2010) Plant-specific glycosylation patterns in the context of therapeutic protein production. Plant Biotechnol J 8(5):564–587
20. Webster DE, Thomas MC (2012) Post-translational modification of plant-made foreign proteins; glycosylation and beyond. Biotechnol Adv 30(2):410–418
21. Ruybicki EP (2009) Plant-produced vaccines: promise and reality. Drug Discov Today 14(1–2):16–24
22. Hellwig S, Drossard J, Twyman RM, Fischer R (2004) Plant cell cultures for the production of recombinant proteins. Nat Biotechnol 22(11):1415–1422
23. Xu J, Ge X, Dolan MC (2011) Towards high-yield production of pharmaceutical proteins with plant cell suspension cultures. Biotechnol Adv 29(3):278–299
24. Buono S, Langellotti AL, Martello A, Rinna F, Fogliano V (2014) Functional ingredients from microalgae. Food Funct 5(8):1669–1685
25. Raja R, Hemaiswarya S, Kumar NA, Sridhar S, Rengasamy R (2008) A perspective on the biotechnological potential of microalgae. Crit Rev Microbiol 34(2):77–88
26. Georgianna DR, Mayfield SP (2012) Exploiting diversity and synthetic biology for the production of algal biofuels. Nature 488(7411):329–335
27. Mata TM, Martins AA, Caetano NS (2010) Microalgae for biodiesel produciton and other applications: A review. Renew Sust Energ Rev 14(1):217–232
28. Moody JW, McGinty CM, Quinn JC (2014) Global evaluation of biofuel potential from microalgae. Proc Natl Acad Sci U S A 111(23):8691–8696
29. Wijffels RH, Barbosa MJ (2010) An outlook on microalgal biofuels. Science 329(5993):796–799
30. Franklin SE, Mayfield SP (2004) Prospects for molecular farming in the green alga Chlamydomonas. Curr Opin Plant Biol 7(2):159–165
31. Rasala BA, Mayfield SP (2015) Photosynthetic biomanufacturing in green algae; production of recombinant proteins for industrial, nutritional, and medical uses. Photosynth Res 123(3):227–239
32. Rosales-Mendoza S, Paz-Maldonado LM, Soria-Guerra RE (2012) Chlamydomonas reinhardtii as a viable platform for the production of recombinant proteins: current status and perspectives. Plant Cell Rep 31(3):479–494
33. Walker TL, Purton S, Becker DK, Collet C (2005) Microalgae as bioreactors. Plant Cell Rep 24(11):629–641
34. Hempel F, Lau J, Klingl A, Maier UG (2011) Algae as protein factories: expression of a human antibody and the respective antigen in the diatom Phaeodactylum tricornutum. PLoS One 6(12):e28424
35. Hempel F, Maier UG (2012) An engineered diatom acting like a plasma cell secreting human IgG antibodies with high efficiency. Microb Cell Factories 11:126
36. Tran M, Zhou B, Pettersson PL, Gonzalez MJ, Mayfield SP (2009) Synthesis and assembly of a

full-length human monoclonal antibody in algal chloroplasts. Biotechnol Bioeng 104(4):663–673

37. Tran M, Van C, Barrera DJ, Pettersson PL, Peinado CD, Bui J, Mayfield SP (2013) Production of unique immunotoxin cancer therapeutics in algal chloroplasts. Proc Natl Acad Sci U S A 110(1):E15–E22
38. Tran M, Henry RE, Siefker D, Van C, Newkirk G, Kim J, Bui J, Mayfield SP (2013) Production of anticancer immunotoxins in algae: ribosome inactivating proteins as fusion partners. Biotechnol Bioeng 110(11):2826–2835
39. Mayfield SP, Franklin SE, Lerner RA (2003) Expression and assembly of a fully active antibody in algae. Proc Natl Acad Sci U S A 100(2):438–442
40. Barrera DJ, Rosenberg JN, Chiu JG, Chang YN, Debatis M, Ngoi SM, Chang JT, Shoemaker CB, Oyler GA, Mayfield SP (2015) Algal chloroplast produced camelid V H antitoxins are capable of neutralizing botulinum neurotoxin. Plant Biotechnol J 13(1):117–124
41. Dreesen IA, Charpin-El Hamri G, Fussenegger M (2010) Heat-stable oral alga-based vaccine protects mice from Staphylococcus aureus infection. J Biotechnol 145(3):273–280
42. Dauvillee D, Delhaye S, Gruyer S, Slomianny C, Moretz SE, d'Hulst C, Long CA, Ball SG, Tomavo S (2010) Engineering the chloroplast targeted malarial vaccine antigens in Chlamydomonas starch granules. PLoS One 5(12):e15424
43. Gregory JA, Li F, Tomosada LM, Cox CJ, Topol AB, Vinetz JM, Mayfield S (2012) Algae-produced Pfs25 elicits antibodies that inhibit malaria transmission. PLoS One 7(5):e37179
44. Gregory JA, Topol AB, Doerner DZ, Mayfield S (2013) Alga-produced cholera toxin-Pfs25 fusion proteins as oral vaccines. Appl Environ Microbiol 79(13):3917–3925
45. Jones CS, Luong T, Hannon M, Tran M, Gregory JA, Shen Z, Briggs SP, Mayfield SP (2013) Heterologous expression of the C-terminal antigenic domain of the malaria vaccine candidate Pfs48/45 in the green algae Chlamydomonas reinhardtii. Appl Microbiol Biotechnol 97(5):1987–1995
46. Li S, Tsai H (2009) Transgenic microalgae as a nonantibiotic bactericide producer to defend against bacterial pathogen infection in the fish digestive tract. Fish Shellfish Immunol 26(2):316–325
47. Feng S, Feng W, Zhao L, Gu H, Li Q, Shi K, Guo S, Zhang N (2014) Preparation of transgenic Dunaliella salina for immunization against white spot syndrome virus in crayfish. Arch Virol 159(3):519–525
48. Geng D, Wang Y, Wang P, Li W, Sun Y (2003) Stable expression of hepatitis B surface antigen gene in Dunaliella salina (Chlorophyta). J Appl Phycol 15(6):451–456
49. Demurtas OC, Massa S, Ferrante P, Venuti A, Franconi R, Giuliano G (2013) A Chlamydomonas-derived Human Papillomavirus 16 E7 vaccine induces specific tumor protection. PLoS One 8(4):e61473
50. Sun M, Qian K, Su N, Chang H, Liu J, Shen G (2003) Foot-and-mouth disease virus VP1 protein fused with cholera toxin B subunit expressed in Chlamydomonas reinhardtii chloroplast. Biotechnol Lett 25(13):1087–1092
51. Surzycki R, Greenham K, Kitayama K, Dibal F, Wagner R, Rochaix JD, Ajam T, Surzycki S (2009) Factors effecting expression of vaccines in microalgae. Biologicals 37(3):133–138
52. He DM, Qian KX, Shen GF, Zhang ZF, Li YN, Su ZL, Shao HB (2007) Recombination and expression of classical swine fever virus (CSFV) structural protein E2 gene in Chlamydomonas reinhardtii chroloplasts. Colloids Surf B: Biointerfaces 55(1):26–30
53. Soria-Guerra RE, Ramírez-Alonso JI, Ibánez-Salazar A, Govea-Alonso DO, Paz-Maldonado LMT, Banuelos-Hernández B, Korban SS, Rosales-Mendoza S (2014) Expression of an HBcAg-based antigen carrying angiotensin II in Chlamydomonas reinhardtii as a candidate hypertension vaccine. Plant Cell Tissue Organ Cult 116(2):133–139
54. Hawkins RL, Nakamura M (1999) Expression of human growth hormone by the eukaryotic alga, Chlorella. Curr Microbiol 38(6):335–341
55. Manuell AL, Beligni MV, Elder JH, Siefker DT, Tran M, Weber A, McDonald TL, Mayfield SP (2007) Robust expression of a bioactive mammalian protein in Chlamydomonas chloroplast. Plant Biotechnol J 5(3):402–412
56. Eichler-Stahlberg A, Weisheit W, Ruecker O, Heitzer M (2009) Strategies to facilitate transgene expression in Chlamydomonas reinhardtii. Planta 229(4):873–883
57. Rasala BA, Muto M, Lee PA, Jager M, Cardoso RM, Behnke CA, Kirk P, Hokanson CA, Crea R, Mendez M, Mayfield SP (2010) Production of therapeutic proteins in algae, analysis of expression of seven human proteins in the chloroplast of Chlamydomonas reinhardtii. Plant Biotechnol J 8(6):719–733
58. Chen Y, Wang Y, Sun Y, Zhang L, Li W (2001) Highly efficient expression of rabbit neutrophil peptide-1 gene in Chlorella ellipsoidea cells. Curr Genet 39(5–6):365–370
59. Bai LL, Yin WB, Chen YH, Niu LL, Sun YR, Zhao SM, Yang FQ, Wang RR, Wu Q, Zhang XQ, Hu ZM (2013) A new strategy to produce a defensin: stable production of mutated NP-1 in nitrate reductase-deficient Chlorella ellipsoidea. PLoS One 8(1):e54966
60. Yang Z, Li Y, Chen F, Li D, Zhang Z, Liu Y, Zheng D, Wang Y, Shen G (2006) Expression of human soluble TRAIL in Chlamydomonas reinhardtii chloroplast. Chin Sci Bull 51(14):1703–1709
61. Wang X, Brandsma M, Tremblay R, Maxwell D, Jevnikar AM, Huner N, Ma S (2008) A novel expression platform for the production of diabetes-

associated autoantigen human glutamic acid decarboxylase (hGAD65). BMC Biotechnol 13(4):460–70

62. Chai X, Chen H, Xu W, Xu Y (2013) Expression of soybean Kunitz trypsin inhibitor gene SKTI in Dunaliella salina. J Appl Phycol 25(1):139–144
63. Sun Y, Yang Z, Gao X, Li Q, Zhang Q, Xu Z (2005) Expression of foreign genes in Dunaliella by electroporation. Mol Biotechnol 30(3):185–192
64. Kim DH, Kim YT, Cho JJ, Bae JH, Hur SB, Hwang I, Choi TJ (2002) Stable integration and functional expression of flounder growth hormone gene in transformed microalga, Chlorella ellipsoidea. Mar Biotechnol (NY) 4(1):63–73
65. Chen HL, Li SS, Huang R, Tsai HJ (2008) Conditional production of a functional fish growth hormone in the transgenic line of nannochloropsis oculata (eustigmatophyceae). J Phycol 44(3):768–776
66. Yoon SM, Kim SY, Li KF, Yoon BH, Choe S, Kuo MM (2011) Transgenic microalgae expressing Escherichia coli AppA phytase as feed additive to reduce phytate excretion in the manure of young broiler chicks. Appl Microbiol Biotechnol 91(3):553–563
67. Rasala BA, Lee PA, Shen Z, Briggs SP, Mendez M, Mayfield SP (2012) Robust expression and secretion of Xylanase1 in Chlamydomonas reinhardtii by fusion to a selection gene and processing with the FMDV 2A peptide. PLoS One 7(8):e43349
68. Georgianna DR, Hannon MJ, Marcuschi M, Shuiqin W, Botsch K, Lewis AJ, Hyun J, Mendez M, Mayfield SP (2013) Production of recombinant enzymes in the marine alga Dunaliella tertiolecta. Algal Res 2(1):2–9
69. Hou Q, Qiu S, Liu Q, Tian J, Hu Z, Ni J (2013) Selenoprotein-transgenic Chlamydomonas reinhardtii. Nutrients 5(3):624–636
70. Han S, Hu Z, Lei A (2008) Expression and function analysis of the metallothionein-like (MT-like) gene from Festuca rubra in Chlamydomonas reinhardtii chloroplast. Sci China C Life Sci 51(12):1076–1081
71. Borovsky D, Powell CR, Dawson WO, Shivprasad S, Lewandowski D, DeBondt HL, DeRanter C, DeLoof A (eds) (1998) Trypsin modulating oostatic factor (TMOF): a new biorational insecticide against mosquitoes. Insects, Chemical Physiological and Environmental Aspects. Wroclaw University Press, Wroclaw
72. Zhang YK, Shen GF, Ru BG (2006) Survival of human metallothionein-2 transplastomic Chlamydomonas reinhardtii to ultraviolet B exposure. Acta Biochim Biophys Sin (Shanghai) 38(3):187–193
73. Kim J, Mayfield SP (1997) Protein disulfide isomerase as a regulator of chloroplast translational activation. Science 278(5345):1954–1957
74. Schroda M (2004) The Chlamydomonas genome reveals its secrets: chaperone genes and the potential roles of their gene products in the chloroplast. Photosynth Res 82(3):221–240
75. Gregory JA, Mayfield SP (2014) Developing inexpensive malaria vaccines from plants and algae. Appl Microbiol Biotechnol 98(5):1983–1990
76. Specht EA, Mayfield SP (2014) Algae-based oral recombinant vaccines. Front Microbiol 5:60
77. Lauersen KJ, Berger H, Mussgnug JH, Kruse O (2013) Efficient recombinant protein production and secretion from nuclear transgenes in Chlamydomonas reinhardtii. J Biotechnol 167(2):101–110
78. Boynton JE, Gillham NW, Harris EH, Hosler JP, Johnson AM, Jones AR, Randolph-Anderson BL, Robertson D, Klein TM, Shark KB et al (1988) Chloroplast transformation in Chlamydomonas with high velocity microprojectiles. Science 240(4858):1534–1538
79. Debuchy R, Purton S, Rochaix JD (1989) The argininosuccinate lyase gene of Chlamydomonas reinhardtii: an important tool for nuclear transformation and for correlating the genetic and molecular maps of the ARG7 locus. EMBO J 8(10):2803–2809
80. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kroger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jezequel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y, Grigoriev IV (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456(7219):239–244
81. Apt KE, Kroth-Pancic PG, Grossman AR (1996) Stable nuclear transformation of the diatom Phaeodactylum tricornutum. Mol Gen Genet 252(5):572–579
82. Miyahara M, Aoi M, Inoue-Kashino N, Kashino Y, Ifuku K (2013) Highly efficient transformation of the diatom Phaeodactylum tricornutum by multi-pulse electroporation. Biosci Biotechnol Biochem 77(4):874–876
83. Zhang C, Hu H (2013) High-efficiency nuclear transformation of the diatom Phaeodactylum tricornutum by electroporation. Mar Genomics 16:63–66
84. Xie WH, Zhu CC, Zhang NS, Li DW, Yang WD, Liu JS, Sathishkumar R, Li HY (2014) Construction of

novel chloroplast expression vector and development of an efficient transformation system for the diatom phaeodactylum tricornutum. Mar Biotechnol (NY) 16(5):538–546

85. Vieler A, Wu G, Tsai CH, Bullard B, Cornish AJ, Harvey C, Reca IB, Thornburg C, Achawanantakun R, Buehl CJ, Campbell MS, Cavalier D, Childs KL, Clark TJ, Deshpande R, Erickson E, Armenia Ferguson A, Handee W, Kong Q, Li X, Liu B, Lundback S, Peng C, Roston RL, Sanjaya SJP, Terbush A, Warakanont J, Zauner S, Farre EM, Hegg EL, Jiang N, Kuo MH, Lu Y, Niyogi KK, Ohlrogge J, Osteryoung KW, Shachar-Hill Y, Sears BB, Sun Y, Takahashi H, Yandell M, Shiu SH, Benning C (2012) Genome, functional gene annotation, and nuclear transformation of the heterokont oleaginous alga Nannochloropsis oceanica CCMP1779. PLoS Genet 8(11):e1003064
86. Kilian O, Benemann CS, Niyogi KK, Vick B (2011) High-efficiency homologous recombination in the oil-producing alga Nannochloropsis sp. Proc Natl Acad Sci U S A 108(52):21265–21269
87. Radakovits R, Jinkerson RE, Fuerstenberg SI, Tae H, Settlage RE, Boore JL, Posewitz MC (2012) Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropis gaditana. Nat Commun 3:686
88. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 306(5693):79–86
89. Poulsen N, Chesley PM, Kröger N (2006) Molecular genetic manipulation of the diatom Thalassiosira pseudonana (Bacillariophyceae). J Phycol 42(5):1059–1065
90. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen CL, Cognat V, Croft MT, Dent R, Dutcher S, Fernandez E, Fukuzawa H, Gonzalez-Ballester D, Gonzalez-Halphen D, Hallmann A, Hanikenne M, Hippler M, Inwood W, Jabbari K, Kalanon M, Kuras R, Lefebvre PA, Lemaire SD, Lobanov AV, Lohr M, Manuell A, Meier I, Mets L, Mittag M, Mittelmeier T, Moroney JV, Moseley J, Napoli C, Nedelcu AM, Niyogi K, Novoselov SV, Paulsen IT, Pazour G, Purton S, Ral JP, Riano-Pachon DM, Riekhof W, Rymarquis L, Schroda M, Stern D, Umen J, Willows R, Wilson N, Zimmer SL, Allmer J, Balk J, Bisova K, Chen CJ, Elias M, Gendler K, Hauser C, Lamb MR, Ledford H, Long JC, Minagawa J, Page MD, Pan J, Pootakham W, Roje S, Rose A, Stahlberg E, Terauchi AM, Yang P, Ball S, Bowler C, Dieckmann CL, Gladyshev VN, Green P, Jorgensen R, Mayfield S, Mueller-Roeber B, Rajamani S, Sayre RT, Brokstein P, Dubchak I, Goodstein D, Hornick L, Huang YW, Jhaveri J, Luo Y, Martinez D, Ngau WC, Otillar B, Poliakov A, Porter A, Szajkowski L, Werner G, Zhou K, Grigoriev IV, Rokhsar DS, Grossman AR (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318(5848):245–250
91. Kindle KL (1990) High-frequency nuclear transformation of Chlamydomonas reinhardtii. Proc Natl Acad Sci U S A 87(3):1228–1232
92. Brown LE, Sprecher SL, Keller LR (1991) Introduction of exogenous DNA into Chlamydomonas reinhardtii by electroporation. Mol Cell Biol 11(4):2328–2332
93. Kumar SV, Misquitta RW, Reddy VS, Rao BJ, Rajamani MV (2004) Genetic transformation of the green alga—Chlamydomonas reinhardtii by Agrobacterium tumefaciens. Plant Sci 166(3):731–738
94. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J, Salamov A, Terry A, Yamada T, Dunigan DD, Grigoriev IV, Claverie JM, Van Etten JL (2010) The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell 22(9):2943–2955
95. Chow KC, Tung WL (1999) Electrotransformation of Chlorella vulgaris. Plant Cell Rep 18(9):778–780
96. Dawson HN, Burlingame R, Cannons AC (1997) Stable transformation of chlorella: rescue of nitrate reductase-deficient mutants with the nitrate reductase gene. Curr Microbiol 35(6):356–362
97. Tan C, Qin S, Zhang Q, Jiang P, Zhao F (2005) Establishment of a micro-particle bombardment transformation system for Dunaliella salina. J Microbiol 43(4):361–365
98. Feng S, Xue L, Liu H, Lu P (2009) Improvement of efficiency of genetic transformation for Dunaliella salina by glass beads method. Mol Biol Rep 36(6):1433–1439
99. Walker TL, Becker DK, Dale JL, Collet C (2005) Towards the development of a nuclear transformation system for Dunaliella tertiolecta. J Appl Phycol 17(4):363–368
100. Gong Y, Hu H, Gao Y, Xu X, Gao H (2011) Microalgae as platforms for production of recombinant proteins and valuable compounds: progress and prospects. J Ind Microbiol Biotechnol 38(12):1879–1890
101. Leon-Banares R, Gonzalez-Ballester D, Galvan A, Fernandez E (2004) Transgenic microalgae as green cell-factories. Trends Biotechnol 22(1):45–52

102. Potvin G, Zhang Z (2010) Strategies for high-level recombinant protein expression in transgenic microalgae: a review. Y. Le Gal, H. O. Halvorson, Publisher: Springer US, Biotechnol Adv 28(6):910–918
103. De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A (2009) Gene silencing in the marine diatom Phaeodactylum tricornutum. Nucleic Acids Res 37(14):e96
104. Schroda M (2006) RNA silencing in Chlamydomonas: mechanisms and tools. Curr Genet 49(2):69–84
105. Sizova I, Greiner A, Awasthi M, Kateriya S, Hegemann P (2013) Nuclear gene targeting in Chlamydomonas using engineered zinc-finger nucleases. Plant J 73(5):873–882
106. Daboussi F, Leduc S, Marechal A, Dubois G, Guyot V, Perez-Michaut C, Amato A, Falciatore A, Juillerat A, Beurdeley M, Voytas DF, Cavarec L, Duchateau P (2014) Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. Nat Commun 5:3831
107. Weyman PD, Beeri K, Lefebvre SC, Rivera J, McCarthy JK, Heuberger AL, Peers G, Allen AE, Dupont CL (2014) Inactivation of phaeodactylum tricornutum urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. Plant Biotechnol J 13(4):460-70
108. Fuhrmann M, Hausherr A, Ferbitz L, Schodl T, Heitzer M, Hegemann P (2004) Monitoring dynamic expression of nuclear genes in Chlamydomonas reinhardtii by using a synthetic luciferase reporter gene. Plant Mol Biol 55(6):869–881
109. Fuhrmann M, Oertel W, Hegemann P (1999) A synthetic gene coding for the green fluorescent protein (GFP) is a versatile reporter in Chlamydomonas reinhardtii. Plant J 19(3):353–361
110. Shao N, Bock R (2008) A codon-optimized luciferase from Gaussia princeps facilitates the in vivo monitoring of gene expression in the model alga Chlamydomonas reinhardtii. Curr Genet 53(6):381–388
111. Hempel F, Bozarth AS, Lindenkamp N, Klingl A, Zauner S, Linne U, Steinbuchel A, Maier UG (2011) Microalgae as bioreactors for bioplastic production. Microb Cell Factories 10:81
112. Mayfield SP, Schultz J (2004) Development of a luciferase reporter gene, luxCt, for Chlamydomonas reinhardtii chloroplast. Plant J 37(3):449–458
113. Gimpel JA, Hyun JS, Schoepp NG, Mayfield SP (2014) Production of recombinant proteins in microalgae at pilot greenhouse scale. Biotechnol Bioeng. doi:10.1002/bit.25357
114. Neupert J, Karcher D, Bock R (2009) Generation of Chlamydomonas strains that efficiently express nuclear transgenes. Plant J 57(6):1140–1150
115. Specht E, Miyake-Stoner S, Mayfield S (2010) Micro-algae come of age as a platform for recombinant protein production. Biotechnol Lett 32(10):1373–1383
116. Poulsen N, Kroger N (2005) A new molecular tool for transgenic diatoms: control of mRNA and protein biosynthesis by an inducible promoter-terminator cassette. FEBS J 272(13):3413–3423
117. Lumbreras V, Stevens DR, Purton S (1998) Efficient foreign gene expression in Chlamydomonas reinhardtii mediated by an endogenous intron. Plant J 14(4):441–447
118. Baiet B, Burel C, Saint-Jean B, Louvet R, Menu-Bouaouiche L, Kiefer-Meyer MC, Mathieu-Rivet E, Lefebvre T, Castel H, Carlier A, Cadoret JP, Lerouge P, Bardor M (2011) N-glycans of Phaeodactylum tricornutum diatom and functional characterization of its N-acetylglucosaminyltransferase I enzyme. J Biol Chem 286(8):6152–6164
119. Mathieu-Rivet E, Kiefer-Meyer MC, Vanier G, Ovide C, Burel C, Lerouge P, Bardor M (2014) Protein N-glycosylation in eukaryotic microalgae and its impact on the production of nuclear expressed biopharmaceuticals. Front Plant Sci 5:359
120. Mathieu-Rivet E, Scholz M, Arias C, Dardelle F, Schulze S, Le Mauff F, Teo G, Hochmal AK, Blanco-Rivero A, Loutelier-Bourhis C, Kiefer-Meyer MC, Fufezan C, Burel C, Lerouge P, Martinez F, Bardor M, Hippler M (2013) Exploring the N-glycosylation pathway in Chlamydomonas reinhardtii unravels novel complex structures. Mol Cell Proteomics 12(11):3160–3183
121. Kerrigan JJ, Xie Q, Ames RS, Lu Q (2011) Production of protein complexes via co-expression. Protein Expr Purif 75(1):1–14
122. Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C (1999) Transformation of nonselectable reporter genes in marine diatoms. Mar Biotechnol (NY) 1(3):239–251
123. Rasala BA, Chao SS, Pier M, Barrera DJ, Mayfield SP (2014) Enhanced genetic tools for engineering multigene traits into green algae. PLoS One 9(4):e94028
124. Shi J, Podola B, Melkonian M (2007) Removal of nitrogen and phosphorus from wastewater using microalgae immobilized on twin layers: an experimental study. J Appl Phycol 19(5):417–423
125. Morris CA, Nicolaus B, Sampson V, Harwood JL, Kille P (1999) Identification and characterization of a recombinant metallothionein protein from a marine alga, Fucus vesiculosus. Biochem J 338(Pt 2):553–560
126. Rajamani S, Siripornadulsil S, Falcao V, Torres M, Colepicolo P, Sayre R (2007) Phycoremediation of heavy metals using transgenic microalgae. Adv Exp Med Biol 616:99–109
127. Cai X, Traina S, Sayre RT (1998) Heavy metal binding properties of wild type and transgenic algae (Chlamydomonas sp.). In: New Developments in Marine Biotechnology. Y. Le Gal, H. O. Halvorson: Springer US pp 189–192
128. Wang B, Lan CQ, Horsman M (2012) Closed photobioreactors for production of microalgal biomasses. Biotechnol Adv 30(4):904–912

Strategies and Methodologies for the Co-expression of Multiple Proteins in Plants

17

Albert Ferrer, Monserrat Arró, David Manzano, and Teresa Altabella

Abstract

The first transgenes were introduced in a plant genome more than 30 years ago. Since then, the capabilities of the plant scientific community to engineer the genome of plants have progressed at an unparalleled speed. Plant genetic engineering has become a central technology that has dramatically incremented our basic knowledge of plant biology and has enabled the translation of this knowledge into a number of increasingly complex and sophisticated biotechnological applications, which in most cases rely on the simultaneous co-expression of multiple recombinant proteins from different origins. To meet the new challenges of modern plant biotechnology, the plant scientific community has developed a vast arsenal of innovative molecular tools and genome engineering strategies. In this chapter we review a variety of tools, technologies, and strategies developed to transfer and simultaneously co-express multiple transgenes and proteins in a plant host. Their potential advantages, disadvantages, and future prospects are also discussed.

Keywords

Plant transformation • Plant biotechnology • Multigene transfer • Multiprotein expression • Chloroplast transformation

A. Ferrer (✉) • M. Arró • D. Manzano
Department of Molecular Genetics, Centre for Research in Agricultural Genomics (CRAG) (CSIC-IRTA-UAB-UB),
Campus UAB, Bellaterra-Cerdanyola del Vallès, 08193 Barcelona, Spain

Department of Biochemistry and Molecular Biology, Faculty of Pharmacy, University of Barcelona, 08028 Barcelona, Spain
e-mail: albertferrer@ub.edu

T. Altabella
Department of Molecular Genetics, Centre for Research in Agricultural Genomics (CRAG) (CSIC-IRTA-UAB-UB),
Campus UAB, Bellaterra-Cerdanyola del Vallès, 08193 Barcelona, Spain

Department of Plant Physiology, Faculty of Pharmacy, University of Barcelona, 08028 Barcelona, Spain

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_17

17.1 Introduction

The first transfer of foreign DNA into a plant genome dates back to 1983, when chimeric bacterial genes conferring resistance to aminoglycoside antibiotics were stably integrated into a plant cell genome [1]. Since then, plant transformation has become a standard and fundamental technology that has boosted progress in basic and applied plant research. The impressive advances in plant genome engineering have allowed the vast amount of basic knowledge acquired in plant biology to be translated into a number of useful biotechnological applications. Hence, plant genetic transformation now permits the enhancement of existing primary and secondary metabolites or the reduction of undesirable ones, the production of new compounds not naturally occurring in plants, the expression of recombinant therapeutic proteins, and the modification of crop plants for better agronomical traits such as pathogen resistance, insect tolerance, enhanced nutritional value, higher yield, and other advantageous characteristics. Achieving these and other challenging biotechnological objectives has been possible in most cases thanks to the development of different methods and strategies allowing the simultaneous expression of multiple recombinant proteins, of plant or non-plant origin, in transgenic plant hosts (Fig. 17.1).

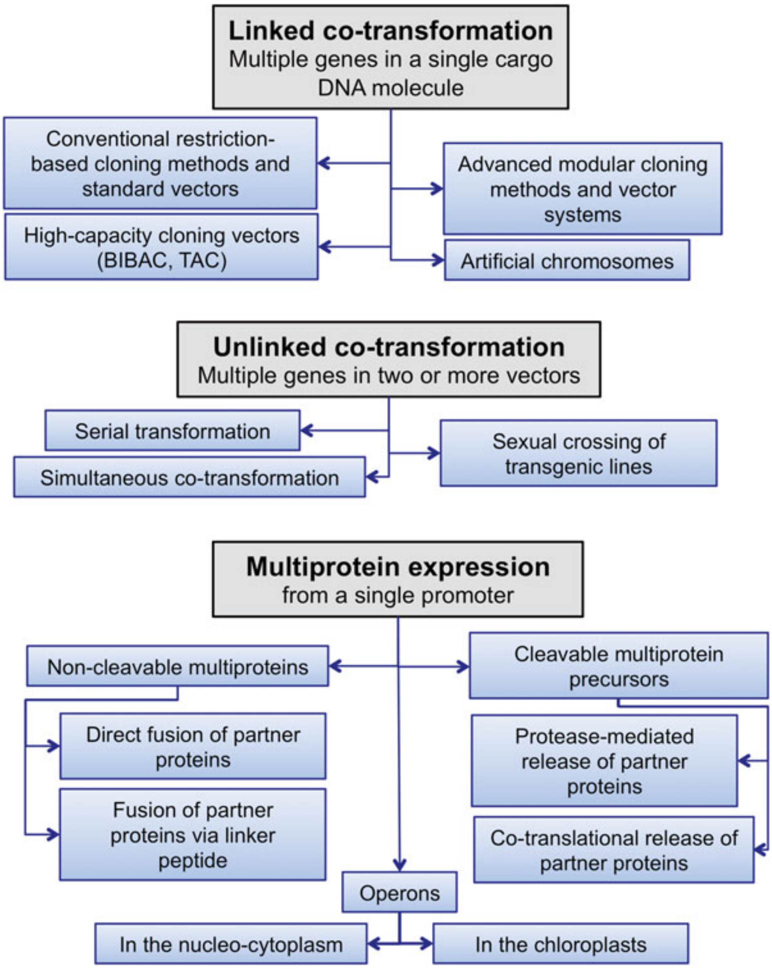

Linked co-transformation
Multiple genes in a single cargo DNA molecule
Conventional restriction-based cloning methods and standard vectors
High-capacity cloning vectors (BIBAC, TAC)
Advanced modular cloning methods and vector systems
Artificial chromosomes
Unlinked co-transformation
Multiple genes in two or more vectors
Serial transformation
Simultaneous co-transformation
Sexual crossing of transgenic lines
Multiprotein expression
from a single promoter
Non-cleavable multiproteins
Direct fusion of partner proteins
Fusion of partner proteins via linker peptide
Cleavable multiprotein precursors
Protease-mediated release of partner proteins
Co-translational release of partner proteins
Operons
In the nucleo-cytoplasm
In the chloroplasts

Fig. 17.1 Strategies for the simultaneous expression of multiple genes and proteins in plants

17.2 Generation of Transgenic Plants

In plant genetic transformation, the foreign DNA harboring the gene of interest (transgene) is first introduced in the recipient cells and subsequently integrated into the nuclear genome in a random manner. Stable DNA integration occurs only in a few cells, whereas in the other cells the DNA is ultimately degraded by endogenous nucleases, although even in this case, the transgene can be expressed for a short time (transient expression) almost immediately after entering the cell. Several methods are currently available for plant cell transformation, although the two most robust, powerful, and widely used approaches are *Agrobacterium tumefaciens*-mediated transformation and microparticle bombardment or biolistics [2–4]. The soil phytopathogenic bacterium *A. tumefaciens* is the causal agent of the neoplastic crown gall disease in a wide range of plant species. *Agrobacterium*-mediated transformation is based on the ability of this natural genetic engineer to deliver a region of the *Agrobacterium* Ti (tumor-inducing) plasmid, the transfer DNA (T-DNA), into the nucleus of plant cells, where it becomes integrated into the genome through a process of illegitimate recombination that has yet to be completely elucidated. The T-DNA is delimited by two 25 bp direct repeat borders that are the only *cis*-acting elements needed for T-DNA transfer [5, 6]. Thus, any DNA placed between the T-DNA borders will be delivered to the host cell using either binary or co-integrative vectors. Whatever the case, the foreign genes to be transferred are placed in the T-DNA region of a disarmed vector (Ti-plasmid) specially suited for this purpose, which is subsequently transformed into the appropriate *Agrobacterium* strain [7]. However, the T-DNA serves only as a cargo vehicle whose mobilization into plant cells will only occur when the *Agrobacterium* strain harboring the engineered Ti plasmid also carries the *Agrobacterium* virulence (*vir*) genes in a separate plasmid (vir helper), in the case of the binary vectors, or in the same plasmid if co-integrative vectors are used. In the binary vectors, which are the most widely used, the *vir*-encoded proteins act in *trans* upon the T-DNA to mediate its processing and subsequent transfer into the plant cell [5, 6]. Apart from bacterial proteins, several host-encoded proteins are also required to complete the transformation process [8].

In addition to stable transformation, the *Agrobacterium*-based system is also used for transient expression of foreign genes (see Chap. 18 for an in-depth discussion on transient expression in plants). This is achieved by flooding the intercellular spaces in the leaves with *Agrobacterium* suspensions harboring the desired recombinant Ti-plasmid, in a process referred to as agroinfiltration [9]. Efficient and robust agroinfiltration methods have been developed for several plant species [9, 10], but *Nicotiana benthamiana* is the most widely used host for transient transgene expression, not only for research purposes [11, 12], but also as a platform for large-scale production of commercially important proteins [13]. Transient expression usually results in very high levels of recombinant protein production from multiple copies of the transgene in a timescale of days, whereas the levels of expression achieved in stable transformation are usually lower because only one or a few copies of the transgene are integrated in the genome. The generation of stably transformed plants is also much more time-consuming and labor-intensive than transient expression approaches. Thus, the possibility of being able to determine the effect of transgene expression in only a matter of days is making transient expression a highly convenient alternative for the rapid assessment of plant-based biotechnological approaches [14, 15]. Nevertheless, it is important to note that the nature of these two expression strategies is essentially different. In addition to considerable differences in the levels of transgene expression, stably transformed plants have to deal with the integrated transgenes and their effects during their entire lifespan and in all tissues if constitutive promoters are used to drive transgene expression. Therefore, results obtained in transient expression experiments may differ from those from stably transformed plants [16, 17].

In contrast to the *Agrobacterium*-mediated transformation method, microparticle bombardment is a direct DNA-transfer technique that relies exclusively on a physical process. Micron-sized tungsten or gold particles (microprojectiles) are coated with DNA and accelerated using commercially available delivery devices into target cells at sufficiently high speed to penetrate the cell wall without killing the cell. Once inside, the DNA is detached from the microprojectiles and eventually becomes stably integrated into the genome through a mechanism that appears to be the same as in *Agrobacterium*-mediated transformed cells [18]. Transformation by particle bombardment is a versatile and effective method because there are no limitations intrinsic to the recipient cell (plant cell type, species or genotype) and the delivery organism (*Agrobacterium*). Thus, biolistics offers the possibility to transform species and genotypes that are not amenable to *Agrobacterium*-mediated transformation. Moreover, biolistics enables the transformation of plastids, which cannot be achieved using *Agrobacterium* because the T-DNA complex is specifically targeted to the nucleus. In fact, both nuclear and plastid genomes can be simultaneously co-transformed using particles coated with a mixture of plastid and nuclear transformation vectors [19]. The latter are standard cloning plasmids only used to propagate the DNA sequence to be transformed in a bacterial host, because no vector sequences are required either for DNA transfer or for integration into the genome of the plant cell [2–4]. On the contrary, since transgene integration into the plastid genome occurs exclusively by homologous recombination, plastid transformation vectors have to include two sequences with homology to the integration site in the plastid genome on either side of the DNA to be transferred [20] (see Sect. 17.4). Particle bombardment is widely used for transient expression studies, but this technique also enables the generation of stably transformed plants due to the possibility of regenerating whole plants from transgenic cells produced by this technology [21–23].

17.3 Multigene Transformation Strategies

In the beginning, the generation of transgenic plants typically involved the transfer of two transgenes into the genome of a model plant: a marker gene for selection and propagation of transformed plants, and the transgene intended to alter the phenotype of plants in a targeted manner. Since then, impressive progress has been made in developing new technologies not only for broadening the range of both model and crop species amenable to transformation, but also to increase the number of foreign genes that can be simultaneously integrated into the genome of the recipient plant. The growing complexity of the new challenges posed by plant basic research and biotechnology has led to a paradigm shift in plant science from "single gene transfer" to "multigene transfer" [24], which has stimulated the development of a new and versatile toolbox for multigene engineering of plant genomes. The opportunity to overcome the limitations of classical plant transformation is currently enabling plant researchers to meet increasingly ambitious and sophisticated biotechnological objectives such as the expression and assembly of macromolecular protein complexes and multimeric proteins [25, 26], the introduction of entire metabolic pathways [27], the assembly of complete synthetic signal transduction pathways [28], and the stacking of multiple agronomic traits [29].

17.3.1 Linked Co-transformation

One of the classical methods used to simultaneously introduce more than one transgene into plants is the stacking in a single cargo DNA fragment of a few expression cassettes (linked co-transformation) arranged in the same transcriptional orientation. These expression cassettes, which usually consist of a promoter, the sequence encoding the protein of interest, and a terminator sequence, are often subjected to a variety of molecular interventions intended to optimize transcription, translation, and protein accumulation in the plant cell environment [30–33]. However, linked co-transformation with conventional binary vectors for *Agrobacterium*-mediated transformation and standard cloning plasmids for biolistic is only suitable for transferring a rather small number of genes [27, 34–36]. The transfer process itself becomes progressively less efficient as the size of the cargo DNA fragment increases. Moreover, the assembly of large fragments in standard binary plasmids and cloning vectors is also a limiting factor due to problems with insert stability and difficulties in finding unique restriction sites during the iterative cloning process [37]. For instance, six of the nine genes introduced into *B. juncea* plants to produce very long chain polyunsaturated fatty acids were assembled together in the same T-DNA using standard restriction enzyme cloning, whereas the remaining three genes were subsequently added using Gateway™ site-specific recombination technology [27].

The limitations concomitant with conventional restriction-based cloning methods, along with the growing need to engineer complex tailor-made multigene expression constructs, has led to the development of a set of advanced modular cloning methods that enable easy and fast assembly of increasingly complex multigene structures from a set of pre-made standard genetic elements or modules. This toolkit includes different promoters, 5′- and 3′-untranslated sequences (5′-UTR and 3′-UTR), transcription terminators, reporter cassettes, gene silencing modules, selectable markers, and sequences coding for antigenic tags and targeting signals, which are usually referred to as "parts" and can be assembled together in multiple combinations following a number of rules known as the "assembly standards" [38]. Modular assembly strategies require much less time and effort than conventional restriction-based cloning methods while providing greater efficiency, versatility, and combinatorial potential, as well as a seemingly limitless reusability, because new composite parts can themselves be included as new parts in the modular assembly pipeline [38]. These advanced DNA assembly methods, which were initially developed and used to implement synthetic biology approaches in microbial systems, can be grouped into those relying on the use of type II restriction endonucleases, such as BioBrick, BglBricks and GoldenGate methods, and those based on sequence-independent overlap techniques, such as Gateway™, circular polymerase extension cloning (CPEC), uracil-specific excision reagent (USER™) cloning, sequence and ligation independent cloning (SLIC) and its commercial version In-Fusion™, seamless ligation cloning extract (SLiCE), Gibson assembly™ [38–40], and precise sequential DNA ligation on solid substrate (PRESSO) [41]. Among them, the Gateway™-based cloning systems are the most widely used to assemble plant multigene expression constructs [42–45], whereas the adaptation of the remaining systems to the specific needs of multigene assembly for plant biotechnology is lagging behind. Even so, modular versions of the GoldenGate system including a large number of plant-specific parts and binary vectors for *Agrobaterium*-mediated transformation have been developed, giving rise to plant-adapted modular cloning systems such as the GoldenBraid [46, 47], the GreenGate [48], and the GoldenGateMoClo [49, 50] methods. In particular, the GoldenGateMoClo system has been successfully used to express in leaves of *N. benthamiana* a 33 kb construct containing 11 transcription units assembled from 44 basic individual modules [49]. Similarly, the PRESSO system has enabled 24 fragments to be assembled in a single expression construct of 17 kb, including seven functional genes involved in ketocarotenoid synthesis, which was introduced into *B. napus* plants via *Agrobacterium*-mediated transformation [51]. A small suit of BioBrick DNA assembly compatible plant transformation vectors has been developed by modifying existing *Agrobacterium* binary plasmids [52], and the capabilities of the USER™ method have been exploited to engineer a collection of binary vectors known as UCE (USER cereal), which performed successfully in *Agrobacterium*-mediated transformation of barley embryos and in the biolistic transformation of barley endosperm cells [53]. The LIC and

SLIC cloning systems have also been adapted to plants with the creation of a set of LIC-compatible plant LIC binary vectors [54], and two SLIC/In-Fusion™-based vectors for biolistics that should provide a basis for developing a further collection of SLIC-based plant transformation vectors [55].

Issues associated with instability of large inserts can be addressed by using non-conventional binary plasmids such as TAC (transformation-competent artificial chromosome) and BIBAC (binary bacterial artificial chromosome) vectors [56, 57]. These extremely low-copy number plasmids combine the capacity to accommodate large fragments of foreign DNA of bacterial artificial chromosomes (BAC) with the specialized features of standard binary plasmids used for *Agrobacterium*-mediated transformation. Moreover, they contain genetic elements added to confer high stability to the recombinant clones in *E. coli* and *Agrobacterium* hosts. Using these vectors, extremely large DNA fragments of up to 150 kb in length have been successfully transferred into plant genomes via *Agrobacterium*-mediated transformation, although the stability of such large inserts is not always guaranteed [58]. The first generation of BAC-based vectors only allowed standard restriction enzyme cloning of the foreign DNA in the T-DNA region [59], but improved versions of these vectors that incorporate recombinase-based cloning systems have been subsequently developed. In spite of this, only a few examples of successful use of these vectors to co-transfer multiple genes have been reported [60, 61].

17.3.2 Unlinked Co-transformation

An alternative to stacking multiple genes in transgenic plants is the use of iterative transformation strategies such as serial transformation (super-transformation), in which the genes of interest are introduced into the same transgenic line through successive rounds of transformation [62, 63] or sexual crossing of transgenic lines carrying different transgenes to bring them together into the same line [64–67]. These unlinked co-transformation approaches may also be combined in an appropriate way to increase the total number of genes introduced [65, 68], but even so they still have some important drawbacks. The different transgenes are integrated in separate loci that can segregate apart in later generations, and both strategies are labor-intensive and time-consuming. Serial transformation may have the additional drawback of requiring a new selectable marker gene for each transformation step, whereas combining genes by sexual crossing cannot be used in crop plants propagated vegetatively [24, 37].

Unlinked multigene transfer can also be achieved by simultaneous transformation of plants with two or more plasmids harboring the genes of interest, in such a way that transgenic plants can be obtained in a single generation. This approach overcomes many of the limitations of conventional multigene transformation methods and is compatible with both *Agrobacterium*-mediated transformation and biolistics, although *Agrobacterium*-mediated unlinked co-transformation is more challenging than co-transformation using biolistic. This is because two compatible binary plasmids need to be maintained in the same bacterial strain or, alternatively, co-infection of the same plant cell is required with two independent strains of *Agrobacterium*, each harboring a different binary vector [69]. Furthermore, multiple T-DNAs tend to integrate in the genome with less efficiency, are prone to disperse to more than one locus, their arrangement may vary depending on the *Agrobacterium* strain, and the ratio of the different transgenes is difficult to control [3, 70]. Unlinked co-transformation using biolistics is more straightforward and powerful. Individual expression cassettes are cloned in separate plasmids, mixed, and loaded onto the metal particles. Upon delivery into the plant cell, all transgenes usually integrate at the same locus as a multigene array, regardless of how many different expression cassettes have been used, so that transgenes are unlikely to segregate apart in subsequent generations [3]. A paradigmatic example of the enormous potential of this multigene transformation approach is the introduction of up to 13 different genes into the rice genome in a single transformation step [21].

The particular features of unlinked co-transformation with biolistic have been exploited to develop the so-called combinatorial transformation strategy [22], which has proven highly successful for metabolic engineering purposes. This innovative approach takes advantage of the inherent variability in the number of transgenes that can be simultaneously integrated using biolistic to create a library of transgenic plants carrying as many random combinations of transgenes as possible in a single generation, instead of undertaking the much more time-consuming and labor-intensive systematic generation of a set of transgenic lines carrying pre-determined combinations of the same transgenes. The subsequent metabolic profiling of the library of transgenic lines allows, on the one hand, dissection and analysis of the pathway and, on the other hand, identification of the most suitable combination of transgenes to achieve the desired metabolic trait and even production of novel metabolites (combinatorial biochemistry) resulting from new combinations of pathway enzymes [22]. Combinatorial transformation has also been successfully applied to simultaneously engineer different metabolic pathways, as demonstrated in maize plants that accumulate high levels of vitamins A, C and B_9 [23].

17.3.3 Genomic Environment and Multigene Expression

A common limitation of the transformation approaches described above is that different transgenes can be expressed at highly variable levels. This is particularly important when equimolar ratios of proteins expressed from different transgenes are required, as can be the case with monomeric constituents of multimeric proteins such as antibodies or enzymes with multiple subunits, and multienzymatic complexes that couple individual enzyme reactions for the channeling of pathway intermediates [71]. Unbalanced gene expression might occur even if the transgenes are physically linked and/or the same or similar promoters are used to drive their expression. In this regard, the impact of repeatedly using the same promoter on the expression of the different transgenes is still a matter of controversy. Using the same promoter may trigger transgene silencing via different molecular mechanisms such as instability by homology-dependent recombination of repeated sequences in the T-DNA prior to or during integration and epigenetic modifications of transgene sequences, particularly *de novo* cytosine methylation, after integration into the host genome. Methylation of the transcribed region has been associated with posttranscriptional gene silencing, whereas methylation of promoter sequences has been related to transcriptional gene silencing [4]. On the contrary, there are examples demonstrating that the same promoter can be successfully used to drive the expression of multiple transgenes without any detrimental effect [23, 51]. Either way, providing each transgene with a different promoter can prevent the potential problems associated with the use of the same regulatory elements. However, as the number of transgenes to be simultaneously expressed grows, it becomes increasingly difficult to identify new promoters for driving proper transgene expression [72]. The use of synthetic promoters could help alleviate such problems. These artificial promoters typically consist of DNA sequences found in endogenous plant promoters, but arranged and condensed in a way not naturally occurring in plants. An additional advantage of synthetic promoters is that they can be specifically designed to fine-tune the expression of transgenes. In fact, some plant synthetic promoters have already been created, but the use of these chimeric regulatory elements in plants is still at a very early stage [39].

The chances of achieving comparable levels of transgene expression and protein production are even smaller when unlinked transgenes are integrated in independent genomic loci due to the so-called position effects that may affect transgene expression differently. When the DNA is integrated in the genome, the nature of the integration site may have a strong influence on the expression of the transgene, which may be affected by the repressive influence exerted by an unfavorable heterochromatic environment or the activity of nearby *cis*-regulatory elements like enhancers or repressors. Position effects can be attenuated by flanking the genes to be transferred with chromatin boundary elements such as insulators or matrix attachment region (MAR) sequences, although the results obtained with these protective elements have been controversial [4, 29]. The potential problems derived from random integration of transgenes, either linked or unlinked, can also be addressed by applying some of the recently developed strategies for site-specific integration of transgenes at pre-defined genomic loci. These approaches are based on the use of engineered site-specific nucleases such as zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), meganucleases (MNs), and clustered regularly interspaced short palindromic repeats (CRISPR-Cas9). These nucleases can create double strand breaks (DSBs) in pre-determined endogenous genomic loci or pre-integrated sites where the foreign DNA can be inserted taking advantage of the endogenous DNA-repair mechanisms. Error-prone non-homologous end-joining is the most frequent repair mechanism in somatic plant cells whereas homologous recombination is a much less common DNA-repair mechanism [73, 74].

17.3.4 Artificial Chromosomes

As described above, the site of transgene integration into the genome of the transformed plant can affect its expression in different ways. However, the opposite also holds true, since the integrity of the recipient genome can also be severely disturbed by transgene insertion, which may lead to disruption of endogenous genes and negative phenotypes. This potential risk is inherent in any transformation method that ends up with the foreign DNA integrated into the host genome, but it does not exist when using other expression approaches not relying on transgene integration into the genome, such as virus-based expression platforms [75, 76] or plant artificial chromosome-assisted transformation [77]. These expression platforms also eliminate the problems associated with position effects. Artificial chromosomes have the added advantage of being able to accommodate very large amounts of foreign DNA. This feature gives them an extraordinary potential to implement more ambitious and technically demanding multigene interventions, such as sophisticated metabolic engineering approaches involving single or even multiple biochemical pathways [78], or the transfer of large DNA fragments harboring entire plant operon-like clusters of co-regulated genes [14]. Thus, although artificial chromosomes are still in the early stages of application in plant basic research and biotechnology, it is becoming generally accepted that this new tool will play a key role in next-generation transgenic approaches.

The minimal requirements of a synthetic plant chromosome vector include a functional centromere, telomeric sequences at the end of each chromosome arm, a piece of genomic DNA containing origins of replication for maintenance and stability during cell division, and a selectable marker gene [77]. Artificial chromosomes were first developed for yeast and mammalian systems using two different approaches known as bottom-up and top-down [79]. The bottom-up strategy involves *de novo* assembly of artificial chromosomes from their basic constituents, but whether this method will also work to assemble plant artificial chromosomes is still a matter of debate [77, 80]. On the contrary, the top-down approach involves engineering of the natural existing chromosomes within a cell to generate artificial chromosome vectors. This technique has already proven useful and fairly robust in engineering endogenous plant chromosomes. So far, two different top-down routes have been used to generate plant artificial chromosomes, the telomere-mediated chromosomal truncation method (telomere seeding), which has been successfully applied to generate artificial chromosomes in different plant species [77], and a very recently developed method that enables generation of artificial ring chromosomes in *Arabidopsis* [81]. Nevertheless, it remains to be established whether this latter technique will also work in other plant species. Despite the impressive progress made over recent years in developing plant artificial chromosomes, some important challenges still need to be addressed for the artificial chromosome-based vectors to become effective and routine multigene plant expression platforms. These include the standardization of efficient methods for artificial chromosome generation, the improvement of translational inheritance, and the introduction of multiple site-specific recombination systems for targeted loading of multiple genes [77, 82]. Moreover, it remains to be elucidated whether artificial chromosomes will adopt the usual chromatin structure and if transgenes in artificial chromosomes will be subjected to the same epigenetic regulatory mechanisms that modulate gene expression in endogenous chromosomes [29].

17.4 Multiprotein Expression from a Single Promoter

A possible way to increase the chances of obtaining comparable levels of co-expressed proteins is to integrate them in a single multiprotein, cleavable or non-cleavable, encoded by a single monocistronic transcript, or to express the different recombinant proteins from a single polycistronic transcript containing multiple open reading frames that are subsequently translated into the individual proteins. A common feature of these multiprotein approaches is that the expression of the chimeric transcriptional units is driven by a single gene promoter (Fig. 17.2).

17.4.1 Non-cleavable Multiprotein Fusions

Non-cleavable multiprotein fusions consist of at least two individual proteins or protein domains that are covalently linked to provide the biological functions of the individual components (Fig. 17.2a). However, in some instances, direct fusion of two or more proteins may be problematic, due to improper folding, proteolysis or unexpected interactions between protein domains that may weaken, or even impair, their biological function. Indeed, steric effects often lead to loss of function of fused proteins. Fusing partner proteins in-frame with a short flexible linker peptide to separate the individual protein components can help to prevent this problem. The amino acid sequence and the length of the linker peptide are critical issues that need to be carefully addressed to avoid any detrimental impact on the performance of the recombinant protein in terms of yield, stability, and biological activity. These two fusion strategies have primarily been used to express chimeric multiproteins specifically devised for herbivorous insect control, including a variety of insecticidal Cry toxins and plant proteins with pesticide functions [83], and less frequently, to express enzyme fusions for metabolic engineering purposes. Some chimeric bifunctional enzymes have been expressed in different plant species [84–87], but the outcome has not always been as successful as hoped. For example, whereas the expression of an artificial fusion of *E. coli* trehalose-6-phosphate synthase and phosphatase enzymes in rice plants resulted in trehalose levels 200-fold higher than those obtained in tobacco plants co-expressing the individual enzymes [85], a fusion of thiolase and reductase enzymes devised for polyhydroxybutyrate biosynthesis in *Arabidopsis* led to lower levels of this polymer than those obtained in plants expressing the enzymes individually [86].

17.4.2 Cleavable Multiprotein Fusions

The different protein components of a multiprotein can also be connected using a linker peptide including a protease recognition site that can be cleaved once inside the plant cell, thus giving rise to individual proteins (Fig. 17.2b). This approach mimics the strategy used by viruses to produce balanced levels of their protein components in the infected cell [88] and by plants that express polypeptides displaying a broad range of inhibitory activities against insect and plant cysteine proteases [89]. The multiprotein precursor can be processed by plant host-encoded (endogenous) proteases or by accessory proteases supplied exogenously, which can be integrated as an additional component of the multiprotein itself (cleavage in *cis*) or expressed from a different

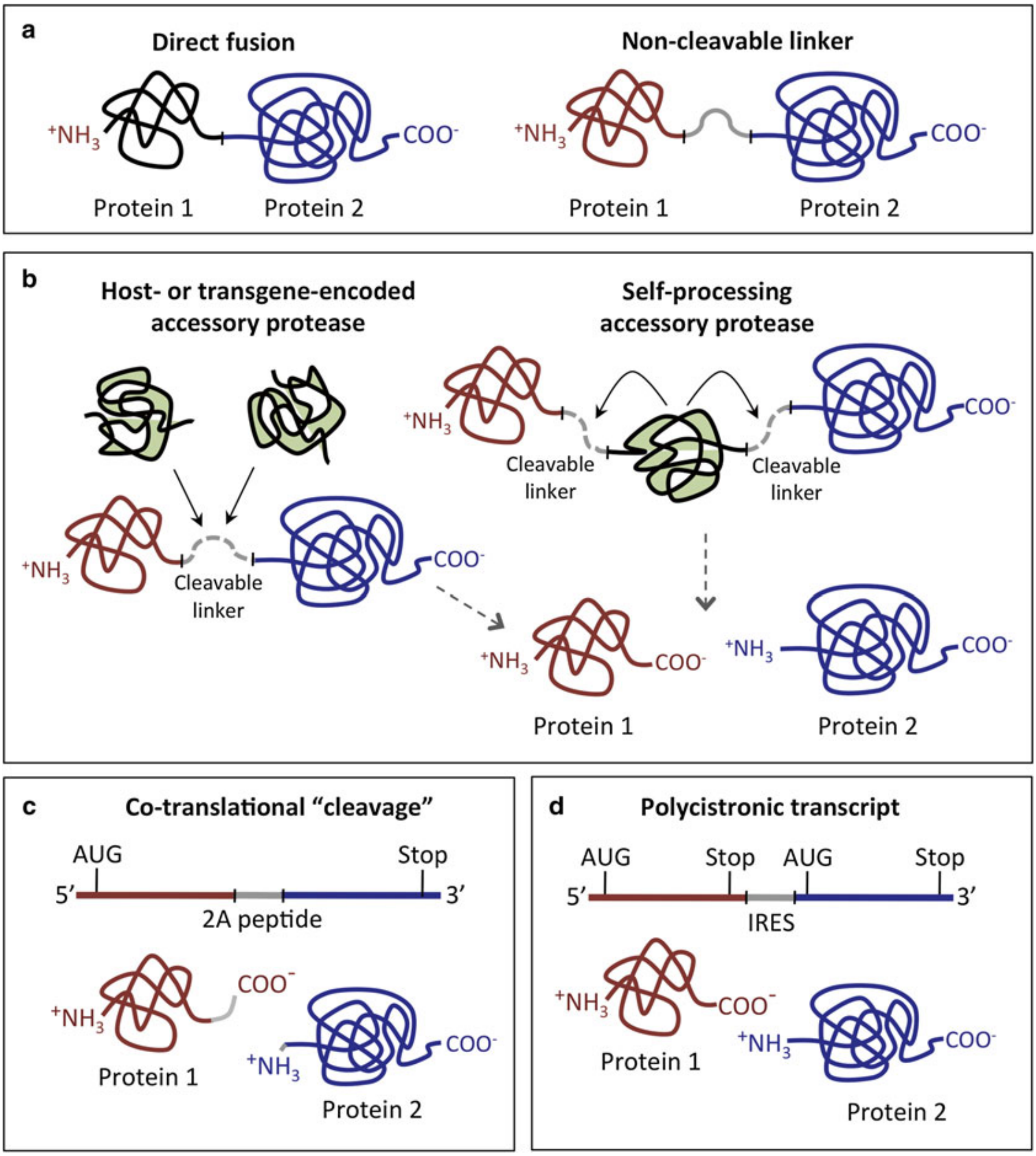


Fig. 17.2 Strategies for the co-expression of multiple proteins from a single gene promoter. Co-expressed proteins can be covalently linked in-frame as a direct fusion or connected in-frame with a non-cleavable short flexible linker peptide (**a**). The linker peptide can also include a protease recognition site (**b**) that can be cleaved by either host-encoded (endogenous) proteases or proteases supplied exogenously from a co-expressed transgene or as a component of the multiprotein itself. Individual proteins can also be co-translationally released from a single transcript containing multiple open reading frames separated by the sequence encoding the foot-and-mouth disease virus (FMDV) 2A peptide (**c**), or individually translated from a polycistronic transcript where the open reading frames are separated by internal ribosome entry sites (IRES) (**d**)

transgene (cleavage in *trans*) [83]. Harnessing endogenous proteases for multiprotein processing does not necessarily mean the precise identity of these proteases is known. Thus, in *Arabidopsis* the expression of chimeric fusions of two protease inhibitors [90, 91] and two antimicrobial proteins [92] gave rise to the individual active proteins after being processed by unidentified endogenous proteases. In other instances, individual proteins are connected with a linker sequence specifically designed to be cleaved by known endogenous protease activities such as deubiquitinating proteases [93] and kex2p-like proteases [94], or proteases supplied exogenously such as the nuclear inclusion a (NIa) protease from plant potyviruses [95]. The latter is the most commonly used protease [96–99], although proteases like the cowpea mosaic virus 24K have also been successfully employed [100]. These accessory proteases are usually included in the multiprotein precursor as a self-processing component flanked by their corresponding peptide target sequences, which, after being cleaved, release the protease that continues processing the multiprotein into their individual protein components. This strategy has been applied in different plant systems to co-express enzymes that introduce a short metabolic pathway [96], transcription factors that upon interaction activate an endogenous metabolic pathway [101], viral capsid protein components [102], and combinations of defensive peptides and proteins [97, 99]. Less frequently, the accessory protease is contributed by an independently expressed transgene [100]. In this case, the expression of transgenes encoding the multiprotein precursor and the *trans*-acting protease must be spatially and temporally coordinated, so that the protease is present and active at the location and/or time of multiprotein expression.

Co-ordinate production of multiple individual proteins can also be achieved by direct co-translational release of individual proteins in host cells without any need for endogenous or exogenous accessory proteases (Fig. 17.2c). This approach relies on the unique properties of the 20-amino acid 2A peptide of foot-and-mouth disease virus (FMDV). The sequence coding for the 2A peptide is included between the sequences coding for the different proteins to be co-expressed. During translation of the engineered transcript, the 2A peptide leads to ribosomal skipping, an alternate mechanism of translation that prevents peptide bond formation between the last two amino acids of the peptide while allowing normal translation to continue. This results in apparent co-translational cleavage of the polyprotein precursor that leads to non-proteolytic dissociation of individual proteins at the C-terminus of the 2A sequence. Co-translational self-dissociation is cell-type independent because it occurs within the ribosome without cytosolic factors [103]. This multiprotein expression system is compatible with both stable and transient expression approaches [104, 105], and has proven functional in different plant systems [104, 106–108]. Moreover, it enables expression of proteins that are to be targeted to different subcellular compartments including the endomembrane system [109] and the chloroplasts [107, 110]. The most basic design of a 2A-based multiprotein cassette is the one allowing production of two individual proteins, although there are successful examples of co-expression of three [104, 105] and even four discrete proteins [111] from a single 2A construct. One of the advantages of this multiprotein expression system is the small size of the 2A peptide compared to the proteases integrated in cleavable multiprotein precursors, which frees up room for the incorporation of additional protein components in the multiprotein precursor. Even so, as in other multiprotein approaches, the number of protein partners that can be integrated in a single 2A multiprotein construct still remains a limiting factor.

The assembly of constructs to express large polycistronic transcripts coding for multiproteins may be hampered by the same practical problems previously described for co-expression of linked genes, namely, difficulties in finding appropriate restriction sites for the cloning process and insert instability. Both co-transformation and serial transformation of plants with more than one 2A multiprotein construct can help to circumvent this limitation, as shown by the introduction of the six enzymes of the *A. thaliana* benzylglucosinolate pathway in tobacco plants. These enzymes were first transiently expressed in *N. benthamiana* through co-infiltration with two different strains of *Agrobacterium* each carrying a single expression construct coding for a set of three pathway enzymes integrated in a 2A multiprotein construct. Once the functionality of the approach was confirmed, the 2A multiprotein constructs were stably transformed into *N. tabacum* plants by applying a serial transformation approach [16]. Nevertheless, co-expression of independent multiprotein transgenes increases the risk of imbalance in the protein production ratios. In fact, expression of a single multiprotein construct does not guarantee absolute stoichiometry levels of the individual proteins either. Final protein production ratios may be affected by the relative position of protein components in the multiprotein construct, since changes of the 2A peptide-flanking context may affect ribosome function and "cleavage" efficiency [104, 106, 108, 111]. On the other hand, the relatively long stretch of 19 amino acid residues from the 2A peptide retained at the C terminus of the proteins, once released from the multiprotein construct, may interfere with their biological activity and/or proper subcellular targeting. The potential adverse effect that this extension may have on the performance of certain proteins has led to the design of strategies to remove these excess amino acid residues. For this, partner proteins are connected with a hybrid linker peptide containing an endogenous protease recognition site fused to the N-terminal end of the 2A peptide. Proteases acting on the N-terminal linker peptide along with ribosome-mediated "cleavage" of the 2A linker results in mature proteins that more closely resemble their native counterparts [106, 112].

17.4.3 Expression of Polycistronic Transcripts in the Nucleo-Cytoplasm

A less widely used approach for protein co-expression involves the use of engineered polycistronic transcripts that incorporate internal ribosome entry sites (IRES) (Fig. 17.2d). These specific RNA motifs were initially discovered in polycistronic viral transcripts and consist of stretches of several hundred nucleotides that form an elaborated secondary structure. This allows eukaryotic ribosomes to bypass the Kozak rule, according to which cytosolic translation in eukaryotes starts at the first AUG codon in a cap-dependent manner. With very few exceptions, downstream AUG codons are not recognized by eukaryotic ribosomes as translation start codons. When IRESs are integrated in polycistronic transcripts, eukaryotic ribosomes are efficiently recruited to initiate translation at internal AUG codons in a cap-independent manner [113]. IRESs have been successfully used to design bicistronic transcripts for protein co-expression in plants [114–117], but the rather moderate efficiency of IRESs-dependent translation compared with cap-dependent translation [115] and 2A multiprotein approaches [116] has limited their widespread use. In general, internal translation initiation is significantly less efficient than translation from the first AUG codon, which may result in severely imbalanced production ratios of individual proteins [116].

Recently, an innovative non-transgenic approach using an expression platform named IL-60 has enabled nucleocytoplasmic expression of an intact bacterial operon in tomato plants [76]. This platform is derived from the geminivirus tomato yellow leaf curl virus (TYLCV) genome, and consists of two components: an expression vector that carries a DNA fragment of the TYLCV containing an origin of replication, a bidirectional promoter and a postulated plant ribosome-binding site, and a helper virus-plasmid vector that promotes cell-to-cell spread of the expression construct throughout the plant [118, 119]. An intact operon encoding four enzymes responsible for the synthesis of the antibiotic compound pyrrolnitrin from tryptophan was transferred from the bacterium *Pseudomonas fluorescens* to the expression vector, and the resulting construct along with the helper construct was subsequently transferred to the plants by root uptake. When using this approach, the introduced trait is not heritable, because the IL-60 recombinant construct including the bacterial operon does not integrate into the plant genome, but expression persists during the entire life span of the plant without causing any disease. The lack of integration also rules out any concern about potential deleterious position effects. Nevertheless, some important questions related to this promising expression platform still require further investigation. These include the elucidation of the mechanism by which cistrons in the unprocessed polycistronic transcript are translated in the cytosol into the four individual enzymes encoded in the operon, the relative levels of expression of the individual enzymes, and finally and perhaps most importantly, whether this novel multigene expression tool can be applied to express other biosynthetic bacterial operons in plants.

17.5 Protein Expression in Plastids

Plastids of higher plants are semi-autonomous organelles that contain their own genome (plastome) and transcription-translation machinery. Transgene expression in plastids is an attractive strategy for the production of recombinant proteins due mainly to their well-recognized capacity to synthesize extraordinary high levels of foreign proteins, which may accumulate to more than 50 % of the total leaf soluble protein [120, 121], and the ease to co-express multiple transgenes arranged in synthetic operons, which has proven particularly useful for metabolic engineering purposes [122, 123]. Other advantages of plastid transformation over nuclear genome transformation are that no transit peptide has to be added to the expressed proteins, which prevents accumulation of non-functional unprocessed precursor forms [124], the precise integration of the transgenes into pre-determined regions of the plastome via homologous recombination [20], the absence of epigenetic effects and gene-silencing mechanisms [125, 126] (see Sect. 17.3.3), and the predominantly maternal inheritance of plastid DNA, which greatly reduces the probability of transgene escape via pollen flow [127, 128].

Plastids arose through endosymbiosis of a cyanobacterial ancestor and have retained numerous prokaryotic features in their genome organization and mechanisms of gene expression, but along evolution, they have also acquired novel non-eubacterial features that make the regulation of plastid gene expression more complex than in prokaryotic organisms. The plastome is a highly polyploid circular double-stranded DNA, which ranges in size from 120 to 220 kb, contains 120–130 genes, and has a quadripartite structure consisting of large and small single copy regions (LSC and SSC) separated by two identical copies of an inverted repeat region (IR_A and IR_B) [20] (Fig. 17.3). The number of plastome copies per plastid and of plastids per cell is dependent on the plant species and cell type, but in all cases, there are a high number of plastome copies per cell, ranging from 500 to 10,000. This is one of the reasons why protein expression in plastids is such an efficient process. The most extreme example is probably the expression in tobacco chloroplasts of lysin, a phage-derived antibiotic protein. Lysin production was so massive (more than 70 % of the plant's total soluble protein) that endogenous plastid protein synthesis was severely compromised [120]. However, despite the structural and functional similarities between the plastome and bacterial genomes, expression of bacterial operons in plastids is not always as straightforward as might be envisaged. Responsible factors may include inefficient recognition of bacterial expression signals by the plastid transcriptional-translational machinery and/or limited stability of the expressed bacterial proteins [123].

17.5.1 Regulation of Gene Expression in Plastids

The regulation of transcription and translation in plastids is reasonably well known. In contrast, the factors affecting the half-life of plastid proteins remain largely unknown. Plastid gene expression is predominantly controlled at the translational level, although some transcriptional regulation also occurs. Reporter gene fusions have identified promoter elements and translation signals in the 5′-UTR and 3′-UTR of plastid mRNAs that are involved in the transcriptional and translational regulation of gene expression [129]. The 5′-UTRs stabilize the mRNAs and facilitate their loading onto the plastid bacterial-

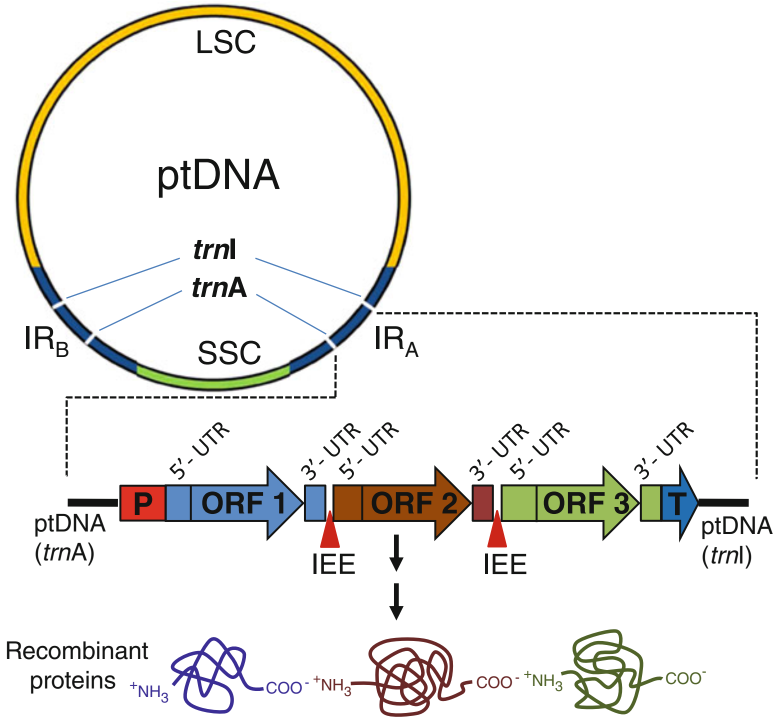


Fig. 17.3 Schematic representation of the plastid genome (ptDNA) harboring a synthetic operon for the co-expression of three recombinant proteins. The quadripartite structure of the ptDNA, including the large and small single copy regions (LSC and SSC), and the two identical copies of an inverted repeat region (IR$_A$ and IR$_B$), which allows duplication of the transgene copy number per genome, is shown. The region between the *trnI* and the *trnA* genes, within the IR region, is the most commonly used site for transgene integration into the ptDNA. The plastid expression construct includes the coding regions of the genes of interest (ORFs 1–3) and different elements involved in efficient operon expression, such as a promoter (P), 5′- and 3′-untranslated sequences (5′-UTR and 3′-UTR), intercistronic expression elements (IEE), and a transcription terminator (T). The plastid transformation construct also contains sequences homologous to the integration site in the ptDNA on either side of the operon to facilitate double integration by homologous recombination

type 70S ribosomes, a process mediated by bacterial-type ribosome-binding sites (RBS), which are a variant of the prokaryotic Shine-Dalgarno sequence found upstream of the translation initiation codon. In some mRNAs with no recognizable Shine-Dalgarno sequences, it seems that mRNA-specific translational activator proteins bind to the 5′-UTR to facilitate translation initiation in a Shine-Dalgarno-independent manner [130, 131]. Although systematic comparative studies are still lacking, it is generally assumed that Shine-Dalgarno-dependent translation rates are overall higher than those that are Shine-Dalgarno-independent. In fact, the strongest translation initiation signals identified so far are found in the 5′-UTR of the *E. coli* phage T7 gene10 (T7g10), which contains a perfect Shine-Dalgarno sequence [132]. Recently, the possibility to enhance translation initiation rates of transgenic mRNA by providing multiple Shine-Dalgarno sequences has been reported, suggesting that a combination of several RBS may be a viable strategy to maximize transgene expression from the plastid genome [133]. The coding sequences immediately downstream of the translation initiation codon also influence plastid recombinant protein production levels. Indeed, certain fusions of the 5′-UTR with sequences encoding the N-terminal coding region, sometimes referred to as 5′ translation control region (5′-TCR), result in a significant enhancement of recombinant protein accumulation. Recent transplantomic studies have demonstrated that major stability determinants of chloroplast proteins are located in the N terminus. Therefore, it seems that rather than stimulating translation, the insertion of specific sequences downstream of the start codon play a role in stabilizing otherwise unstable recombinant proteins. Thus, manipulation of the N terminus of poorly stable recombinant proteins or fusing them to the N terminus of a stable protein can help to solve the stability problem [134]. However, it seems unlikely that protein stability in plastids may be determined only by their N termini. Internal sequence motifs or improper protein folding may also trigger rapid protein degradation but, unfortunately, very little is known about these structural determinants. Thus, rendering more predictable the stability of recombinant proteins in plastids and providing guidelines for stabilizing labile proteins represent major challenges for future research on plastid transgene expression [134]. The 3′-UTRs of plastid mRNAs are also important for mRNA stability but the impact on expression levels is rather limited [130, 135].

Transcription of the plastid genome relies on a plastid-encoded bacterial-type RNA polymerase (PEP) and one or two (depending on the plant species) nucleus-encoded bacteriophage-type RNA polymerases (NEP) [136] that recognize different types of promoters. In general, PEP-type promoters are stronger than NEP promoters. Some promoters of non-plastid origin have also been tested for their capacity to drive transgene expression in plastids, but unless recognized by the chloroplast transcriptional machinery, they have been found to be significantly weaker than the PEP promoters. Most biotechnological approaches utilize the PEP promoter of the *rrn* ribosomal RNA operon (*Prrn*), which is the strongest plastid promoter. However, as the ribosomal RNAs are not translated, *Prrn* needs to be fused with an appropriate translation control sequence (5′-UTR/5′-TCR) to achieve high-level protein accumulation. The choice of the proper translation control signal is extremely important since it may drastically affect protein production yields from the same *Prrn* promoter in a 10,000-fold range [130, 137]. As mentioned above, the most efficient translation control sequences derive from the T7g10 gene [132], which once fused to the *Prrn* promoter resulted in the highest protein accumulation levels reported to date [120]. Strong translation control signals are also found in the 5′-UTR of the *Bacillus thuringiensis cry9Aa2* gene [138] and the plastid *rbcL* gene [139]. The design of chimeric expression elements containing different promoters and 5′-UTRs has also proven to be an appropriate strategy to boost foreign protein accumulation in non-green plastids, such as fruit chromoplasts, tuber amyloplasts, and seed elaioplasts [129, 140].

17.5.2 Expression of Polycistronic Transcripts in Plastids

Chloroplasts are the compartment of the cell where more primary and specialized metabolism takes place, which has made these organelles an attractive target for metabolic engineering purposes, including the implementation of novel metabolic pathways, which often requires the introduction and co-expression of several genes [134]. As described above, the genes in the plastome still retain many elements of the prokaryotic expression machinery and most of them are organized in bacterial-type operons leading to polycistronic transcription units similar to bacterial polycistronic transcripts. Therefore, plastid expression of both entire bacterial operons and synthetic operons containing multiple ORFs from different origins would seem feasible. Indeed, operon expression has been successful in some cases, but there are also examples of poor expression of at least some of the transgenes in the operon [131]. The redesign of bacterial operons by replacing bacterial expression signals with plastid-specific 5′-TCR sequences can improve the expression efficiency of transgenes stacked in operons, as in the case of tobacco plants expressing the fully functional luciferase pathway of the bioluminescent bacterium *Photobacterium leiognathi* in chloroplasts. The resulting autoluminiscent plants, which emit light visible to the naked eye, were created by expressing the six genes of the lux operon (*luxCDABEG*) in the plastid genome under the control of the *Prrn* promoter. Expression of the first pathway enzyme (luxC) was controlled by the 5′-TCR of the *rbcL* gene, whereas the remaining enzymes were expressed from their native translation control sequences [141].

The differences in transcript maturation between bacteria and plastids are a potential problem for efficient operon expression in plastids. Bacterial polycistronic transcripts directly enter translation, while the general consensus is that plastidial ones undergo post-transcriptional processing into mono- or oligocistronic units before translation [142]. This post-transcriptional processing step can be essential for efficient translation, particularly of the downstream cistrons [143, 144], but not all polycistronic transcripts in plastids need intercistronic processing to be translated. Whether or not intercistronic processing is required for efficient transgene expression from operons is currently unpredictable and needs to be determined case by case. Expression of the complete *B. thuringensis cry*2Aa2 operon and six more artificial operons in *N. tabacum* chloroplasts resulted in the production of large amounts of the foreign proteins, which were predominantly translated from polycistronic transcripts [145, 146]. An artificial metabolic operon containing the genes coding for the enzymes β-carotene ketolase and β-carotene hydroxylase from a marine bacterium *Brevundimonas* sp. has also been successfully expressed from unprocessed polycistronic transcripts in transplantomic tobacco plants to produce astaxanthin [147]. The carotenoid pigment levels in the transplantomic plants were far higher than those in transgenic tobacco plants expressing the same enzymes from the nuclear genome, where only traces of this pigment were detected [107], thereby illustrating the advantages of metabolic engineering based on transplantomic approaches rather than on nuclear transformation. However, in other instances transgene expression in plastids has been rather poor or even unsuccessful as a result of inefficient translation of polycistronic mRNAs [148, 149]. Transplantomic experiments to analyze RNA processing in the plastidial *psbB* operon have identified a minimum sequence element known as the intercistronic expression element (IEE), which proved necessary and sufficient to trigger intercistronic processing when placed in a chimeric context [150]. This allows the introduction of processing signals into synthetic chloroplast operons, thereby minimizing the risk of inefficient translation. The functionality of the IEE element has been demonstrated in transplastomic tobacco and tomato plants engineered to produce high levels of vitamin E. When a synthetic operon consisting of the three genes encoding the key enzymes of this biosynthetic pathway, namely homogentisate phytyl-transferase, tocopherol cyclase, and γ-tocopherol methyl-transferase, separated by IEE sequences, was expressed in the transplantomic plants, the vitamin E levels were significantly higher than those in plants expressing an equivalent operon where no IEEs were included [151]. Thus, the IEE provides a valuable sequence context-independent tool for synthetic operon design that increases the chances of successful operon expression in transgenic plastids and opens the possibility to express much larger operons encoding more complex metabolic pathways in engineered plastid genomes.

The integration of foreign DNA into the plastid genome occurs exclusively via homologous recombination mediated by a RecA-type system [152]. This particular feature enables targeted integration of the transgenes, thus eliminating any concern about position effects and negative phenotypes resulting from disruption of endogenous genes (see Sect. 17.3.3). The high activity of this plastid recombination system also facilitates the simultaneous integration of the transgenes in two distinct regions of the plastid genome by co-transformation experiments using microprojectiles loaded with two or more plasmids [153]. The number of insertion sites may affect the level of protein accumulation, since the transgene copy number per genome is duplicated if the integration occurs in the IR region rather than in a single copy region. The level of translatable mRNA and protein can be further increased if the transgene is inserted between genes of a heavily transcribed operon [145]. The most commonly used site of integration is the transcriptionally active intergenic region between the *trn*I (tRNA-Ile) and *trn*A (tRNA-Ala) genes, within the *rrn* operon, located in the IR regions of the plastome (Fig. 17.3). It appears that this site allows highly efficient transgene integration and expression [121, 125, 127, 145, 154]. In fact, the expression of the *lux* operon integrated at this site is 25-fold higher than when insertion occurs into a transcriptionally silent spacer region (*rps12*-*trnV* locus) [141].

So far, biolistics has been the method of choice for plastid transformation (see Sect. 17.2), although polyethylene glycol (PEG)-mediated protoplast transformation is occasionally used as an alternative method [130, 137]. Plastid transformation vectors are based on standard *E. coli* plasmids that have been engineered to carry two sequences of about 1 kb in length, homologous to the integration site in the plastome, flanking the cloning site. Similarly to the toolkit for nuclear genome engineering (see Sect. 17.3.1), a large set of genetic elements for plastid genome engineering has been identified and is currently available. This toolkit contains selectable marker genes, reporter genes, transcriptional and translational regulatory sequences such as promoters, 5′-regulatory sequences (5′-UTR and 5′-TCR), and 3′-regulatory regions (3′-UTR), synthetic sequences for homologous recombination devoid of undesirable restriction sites, and synthetic elements for expression in plastids of non-green tissues [123, 129, 130, 134, 137, 140, 155, 156]. Thus, the sequences coding for the proteins of interest can be easily combined with these modular elements to create recombinant plasmids for expression of tailor-made transgenes in the plastids, either single genes or multiple genes arranged in operons (Fig. 17.3).

In spite of the enormous progress in plastid genome engineering and recombinant protein expression in recent years, one of the most important limitations that still remains to be overcome concerns the small number of plant species amenable to plastid transformation, monocots being the most recalcitrant [130, 134]. Thus, developing effective and robust plastid transformation protocols for important monocot crops still represents a formidable challenge to be addressed in plant biotechnology.

Acknowledgments We gratefully acknowledge the support of the Spanish Ministerio de Economía y Competitividad (AGL2013-43522-R) and the Generalitat de Catalunya (2014 SGR 1434).

References

1. Fraley RT, Rogers SG, Horsch RB, Sanders PR, Flick JS, Adams SP, Bittner ML, Brand LA, Fink CL, Fry JS, Galluppi GR, Goldberg SB, Hoffmann NL, Woo SC (1983) Expression of bacterial genes in plant cells. Proc Natl Acad Sci U S A 80:4803–4807
2. Lorence A, Verpoorte R (2004) Gene transfer and expression in plants. Methods Mol Biol 267:329–350
3. Altpeter F, Baisakh N, Beachy R, Bock R, Capell T, Christou P, Daniell H, Datta K, Datta S, Dix PJ, Fauquet C, Huang N, Kohli A, Mooibroek H, Nicholson L, Nguyen TT, Nugent G, Raemakers K, Romano A, Somers DA, Stoger E, Taylor N, Visser R (2005) Particle bombardment and the genetic enhancement of crops: myths and realities. Mol Breed 15:305–327
4. Kohli A, Miro B, Twyman RM (2010) Transgene integration, expression and stability in plants: strategies for improvements. In: Transgenic crop plants. Springer, Berlin, pp 201–237
5. Pitzschke A, Hirt H (2010) New insights into an old story: Agrobacterium-induced tumour formation in plants by plant transformation. EMBO J 29:1021–1032
6. Păcurar DI, Thordal-Christensen H, Păcurar ML, Pamfil D, Botez C, Bellini C (2011) Agrobacterium tumefaciens: from crown gall tumors to genetic transformation. Physiol Mol Plant Pathol 76:76–81
7. Lee LY, Gelvin SB (2008) T-DNA Binary vectors and systems. Plant Physiol 146:325–332
8. Gelvin SB (2010) Plant proteins involved in Agrobacterium-mediated genetic transformation. Annu Rev Phytopathol 48:45–68
9. Kapila J, De Rycke R, Van Montagu M, Angenon G (1997) An Agrobacterium-mediated transient gene expression system for intact leaves. Plant Sci 122:101–108
10. Wroblewski T, Tomczak A, Michelmore R (2005) Optimization of Agrobacterium-mediated transient assays of gene expression in lettuce, tomato and Arabidopsis. Plant Biotechnol J 3:259–273
11. Wood CC, Petrie JR, Shrestha P, Mansour MP, Nichols PD, Green AG, Singh SP (2009) A leaf-based assay using interchangeable design principles to rapidly assemble multistep recombinant pathways. Plant Biotechnol J 7:914–924

12. Liu Q, Manzano D, Tanić N, Pesic M, Bankovic J, Pateraki I, Ricard L, Ferrer A, De Vos R, van de Krol S, Bouwmeester H (2014) Elucidation and in planta reconstitution of the parthenolide biosynthetic pathway. Metab Eng 23:145–153
13. Chen Q, Lai H, Hurtado J, Stahnke J (2013) Agroinfiltration as an effective and scalable strategy of gene delivery for production of pharmaceutical proteins. Adv Tech Biol Med 1:103
14. Nützmann H-W, Osbourn A (2014) Gene clustering in plant specialized metabolism. Curr Opin Biotechnol 26:91–99
15. Sainsbury F, Lomonossoff GP (2014) Transient expressions of synthetic biology in plants. Curr Opin Plant Biol 19:1–7
16. Conley AJ, Zhu H, Le LC, Jevnikar AM, Lee BH, Brandle JE, Menassa R (2011) Recombinant protein production in a variety of Nicotiana hosts: a comparative analysis. Plant Biotechnol J 9:434–444
17. Møldrup ME, Geu-Flores F, de Vos M, Olsen CE, Sun J, Jander G, Halkier BA (2012) Engineering of benzylglucosinolate in tobacco provides proof-of-concept for dead-end trap crops genetically modified to attract Plutella xylostella (diamondback moth). Plant Biotechnol J 10:435–442
18. Somers DA, Makarevitch I (2004) Transgene integration in plants: poking or patching holes in promiscuous genomes? Curr Opin Biotechnol 15:126–131
19. Elghabi Z, Ruf S, Bock R (2011) Biolistic co-transformation of the nuclear and plastid genomes. Plant J 67:941–948
20. Verma D, Daniell H (2007) Chloroplast vector systems for biotechnology applications. Plant Physiol 145:1129–1143
21. Chen L, Marmey P, Taylor NJ, Brizard J-P, Espinoza C, D'Cruz P, Huet H, Zhang S, de Kochko A, Beachy RN, Fauquet CM (1998) Expression and inheritance of multiple transgenes in rice plants. Nat Biotechnol 16:1060–1064
22. Zhu C, Naqvi S, Breitenbach J, Sandmann G, Christou P, Capell T (2008) Combinatorial genetic transformation generates a library of metabolic phenotypes for the carotenoid pathway in maize. Proc Natl Acad Sci U S A 105:18232–18237
23. Naqvi S, Zhu C, Farré G, Ramessar K, Bassie L, Breitenbach J, Perez Conesa D, Ros G, Sandmann G, Capell T, Christou P (2009) Transgenic multivitamin corn through biofortification of endosperm with three vitamins representing three distinct metabolic pathways. Proc Natl Acad Sci U S A 106:7762–7767
24. Halpin C (2005) Gene stacking in transgenic plants – the challenge for 21st century plant biotechnology. Plant Biotechnol J 3:141–155
25. Nicholson L, Gonzalez-Melendi P, Van Dolleweerd C, Tuck H, Perrin Y, Ma JKC, Fischer R, Christou P, Stoger E (2004) A recombinant multimeric immunoglobulin expressed in rice shows assembly-dependent subcellular localization in endosperm cells. Plant Biotechnol J 3:115–127
26. Thuenemann EC, Meyers AE, Verwey J, Rybicki EP, Lomonossoff GP (2013) A method for rapid production of heteromultimeric protein complexes in plants: assembly of protective bluetongue virus-like particles. Plant Biotechnol J 11:839–846
27. Wu G, Truksa M, Datla N, Vrinten P, Bauer J, Zank T, Cirpus P, Heinz E, Qiu X (2005) Stepwise engineering to produce high yields of very long-chain polyunsaturated fatty acids in plants. Nat Biotechnol 23:1013–1017
28. Antunes MS, Morey KJ, Smith JJ, Albrecht KD, Bowen TA, Zdunek JK, Troupe JF, Cuneo MJ, Webb CT, Hellinga HW, Medford JI (2011) Programmable ligand detection system in plants through a synthetic signal transduction pathway. PLoS ONE 6, e16292
29. Que Q, Chilton M-DM, de Fontes CM, He C, Nuccio M, Zhu T, Wu Y, Chen JS, Shi L (2010) Trait stacking in transgenic crops: challenges and opportunities. GM Crops 1:220–229
30. Streatfield SJ (2007) Approaches to achieve high-level heterologous protein production in plants. Plant Biotechnol J 5:2–15
31. Egelkrout E, Rajan V, Howard JA (2012) Overproduction of recombinant proteins in plants. Plant Sci 184:83–101
32. Jackson MA, Sternes PR, Mudge SR, Graham MW, Birch RG (2014) Design rules for efficient transgene expression in plants. Plant Biotechnol J 12:925–933
33. Meshcheriakova YA, Saxena P, Lomonossoff GP (2014) Fine-tuning levels of heterologous gene expression in plants by orthogonal variation of the untranslated regions of a nonreplicating transient expression system. Plant Biotechnol J 12:718–727
34. Slater S, Mitsky TA, Houmiel KL, Hao M, Reiser SE, Taylor NB, Tran M, Valentin HE, Rodriguez DJ, Stone DA, Padgette SR, Kishore G, Gruys KJ (1999) Metabolic engineering of Arabidopsis and Brassica for poly(3-hydroxybutyrate-co-3-hydroxyvalerate) copolymer production. Nat Biotechnol 17:1011–1016
35. Bohmert K, Balbo I, Kopka J, Mittendorf V, Nawrath C, Poirier Y, Tischendorf G, Trethewey RN, Willmitzer L (2000) Transgenic Arabidopsis plants can accumulate polyhydroxybutyrate to up to 4% of their fresh weight. Planta 211:841–845
36. Huang JC, Zhong YJ, Liu J, Sandmann G, Chen F (2013) Metabolic engineering of tomato for high-yield production of astaxanthin. Metab Eng 17:59–67
37. Naqvi S, Farré G, Sanahuja G, Capell T, Zhu C, Christou P (2010) When more is better: multigene engineering in plants. Trends Plant Sci 15:48–56
38. Ellis T, Adie T, Baldwin GS (2011) DNA assembly for synthetic biology: from parts to pathways and beyond. Integr Biol 3:109–118
39. Liu W, Yuan JS, Stewart CN (2013) Advanced genetic tools for plant biotechnology. Nat Rev Genet 14:781–793

40. Patron NJ (2014) DNA assembly for plant biology: techniques and tools. Curr Opin Plant Biol 19:14–19
41. Takita E, Kohda K, Tomatsu H, Hanano S, Moriya K, Hosouchi T, Sakurai N, Suzuki H, Shinmyo A, Shibata D (2013) Precise sequential DNA ligation on a solid substrate: solid-based rapid sequential ligation of multiple DNA molecules. DNA Res 20:583–592
42. Chen Q-J, Zhou H-M, Chen J, Wang X-C (2006) A Gateway-based platform for multigene plant transformation. Plant Mol Biol 62:927–936
43. Karimi M, Depicker A, Hilson P (2007) Recombinational cloning with plant Gateway vectors. Plant Physiol 145:1144–1154
44. Untergasser A, Bijl GJM, Liu W, Bisseling T, Schaart JG, Geurts R (2012) One-step Agrobacterium mediated transformation of eight genes essential for rhizobium symbiotic signaling using the novel binary vector system pHUGE. PLoS ONE 7, e47885
45. Vemanna RS, Chandrashekar BK, Hanumantha Rao HM, Sathyanarayanagupta SK, Sarangi KS, Nataraja KN, Udayakumar M (2012) A modified MultiSite Gateway cloning strategy for consolidation of genes in plants. Mol Biotechnol 53:129–138
46. Sarrion-Perdigones A, Falconi EE, Zandalinas SI, Juárez P, Fernández-del-Carmen A, Granell A, Orzaez D (2011) GoldenBraid: an iterative cloning system for standardized assembly of reusable genetic modules. PLoS ONE 6, e21622
47. Sarrion-Perdigones A, Vazquez-Vilar M, Palaci J, Castelijns B, Forment J, Ziarsolo P, Blanca J, Granell A, Orzaez D (2013) GoldenBraid 2.0: a comprehensive DNA assembly framework for plant synthetic biology. Plant Physiol 162:1618–1631
48. Lampropoulos A, Sutikovic Z, Wenzl C, Maegele I, Lohmann JU, Forner J (2013) GreenGate – a novel, versatile, and efficient cloning system for plant transgenesis. PLoS ONE 8, e83043
49. Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S (2011) A modular cloning system for standardized assembly of multigene constructs. PLoS ONE 6, e16765
50. Engler C, Youles M, Gruetzner R, Ehnert T-M, Werner S, Jones JDG, Patron NJ, Marillonnet S (2014) A Golden Gate modular cloning toolbox for plants. ACS Synth Biol. doi:10.1021/sb4001504
51. Fujisawa M, Takita E, Harada H, Sakurai N, Suzuki H, Ohyama K, Shibata D, Misawa N (2009) Pathway engineering of Brassica napus seeds using multiple key enzyme genes involved in ketocarotenoid formation. J Exp Bot 60:1319–1332
52. Boyle PM, Burrill DR, Inniss MC, Agapakis CM, Deardon A, Dewerd JG, Gedeon MA, Quinn JY, Paull ML, Raman AM, Theilmann MR, Wang L, Winn JC, Medvedik O, Schellenberg K, Haynes KA, Viel A, Brenner TJ, Church GM, Shah JV, Silver PA (2012) A BioBrick compatible strategy for genetic modification of plants. J Biol Eng 6:8
53. Hebelstrup KH, Christiansen MW, Carciofi M, Tauris B, Brinch-Pedersen H, Holm PB (2010) UCE: a uracil excision (USER™)-based toolbox for transformation of cereals. Plant Methods 6:1–10
54. De Rybel B, van den Berg W, Lokerse AS, Liao CY, van Mourik H, Moller B, Llavata-Peris CI, Weijers D (2011) A versatile set of ligation-independent cloning vectors for functional studies in plants. Plant Physiol 156:1292–1299
55. Kronbak R, Ingvardsen CR, Madsen CK, Gregersen PL (2014) A novel approach to the generation of seamless constructs for plant transformation. Plant Methods 10:1–10
56. Hamilton CM (1997) A binary-BAC system for plant transformation with high-molecular-weight DNA. Gene 200:107–116
57. Liu YG, Shirano Y, Fukaki H, Yanai Y, Tasaka M, Tabata S, Shibata D (1999) Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. Proc Natl Acad Sci U S A 96:6535–6540
58. Song J, Bradeen JM, Naess SK, Helgeson JP, Jiang J (2003) BIBAC and TAC clones containing potato genomic DNA fragments larger than 100 kb are not stable in Agrobacterium. Theor Appl Genet 107:958–964
59. Shibata D, Liu YG (2000) Agrobacterium-mediated plant transformation with large DNA fragments. Trends Plant Sci 5:354–357
60. Lin L, Liu Y-G, Xu X, Li B (2003) Efficient linking and transfer of multiple genes by a multigene assembly and transformation vector system. Proc Natl Acad Sci U S A 100:5962–5967
61. Buntru M, Gärtner S, Staib L, Kreuzaler F, Schlaich N (2013) Delivery of multiple transgenes to plant cells by an improved version of MultiRound Gateway technology. Transgenic Res 22:153–167
62. Jobling SA, Westcott RJ, Tayal A, Jeffcoat R, Schwall GP (2002) Production of a freeze-thaw-stable potato starch by antisense inhibition of three starch synthase genes. Nat Biotechnol 20:295–299
63. Qi B, Fraser T, Mugford S, Dobson G, Sayanova O, Butler J, Napier JA, Stobart AK, Lazarus CM (2004) Production of very long chain polyunsaturated omega-3 and omega-6 fatty acids in plants. Nat Biotechnol 22:739–745
64. Ma JKC, Hiatt A, Hein M, Vine ND, Wang F, Stabila P, Van Dolleweerd C, Mostov K, Lehner T (1995) Generation and assembly of secretory antibodies in plants. Science 268:716–719
65. Datta K, Baisakh N, Thet KM, Tu J, Datta S (2002) Pyramiding transgenes for multiple resistance in rice against bacterial blight, yellow stem borer and sheath blight. Theor Appl Genet 106:1–8
66. Houshyani B, Assareh M, Busquets A, Ferrer A, Bouwmeester HJ, Kappers IF (2013) Three-step pathway engineering results in more incidence rate and higher emission of nerolidol and improved

attraction of Diadegma semiclausum. Metab Eng 15:88–97

67. van Erp H, Kelly AA, Menard G, Eastmond PJ (2014) Multigene engineering of triacylglycerol metabolism boosts seed oil content in Arabidopsis. Plant Physiol 165:30–36
68. Kebeish R, Niessen M, Thiruveedhi K, Bari R, Hirsch H-J, Rosenkranz R, Stäbler N, Schönfeld B, Kreuzaler F, Peterhänsel C (2007) Chloroplastic photorespiratory bypass increases photosynthesis and biomass production in Arabidopsis thaliana. Nat Biotechnol 25:593–599
69. Ebinuma H, Sugita K, Matsunaga E, Endo S, Yamada K, Komamine A (2001) Systems for the removal of a selection marker and their combination with a positive marker. Plant Cell Rep 20:383–392
70. Kohli A, Twyman RM, Abranches R, Wegel E, Stoger E, Christou P (2003) Transgene integration, organization and interaction in plants. Plant Mol Biol 52:247–258
71. Conrado RJ, Varner JD, DeLisa MP (2008) Engineering the spatial organization of metabolic enzymes: mimicking nature's synergy. Curr Opin Biotechnol 19:492–499
72. Peremarti A, Twyman RM, Gómez-Galera S, Naqvi S, Farré G, Sabalza M, Miralpeix B, Dashevskaya S, Yuan D, Ramessar K, Christou P, Zhu C, Bassie L, Capell T (2010) Promoter diversity in multigene transformation. Plant Mol Biol 73:363–378
73. Podevin N, Davies HV, Hartung F, Nogué F, Casacuberta JM (2013) Site-directed nucleases: a paradigm shift in predictable, knowledge-based plant breeding. Trends Biotechnol 31:375–383
74. Kathiria P, Eudes F (2014) Nucleases for genome editing in crops. Biocatalysis Agric Biotechnol 3:14–19
75. Marton I, Zuker A, Shklarman E, Zeevi V, Tovkach A, Roffe S, Ovadis M, Tzfira T, Vainstein A (2010) Nontransgenic genome modification in plant cells. Plant Physiol 154:1079–1087
76. Mozes-Koch R, Gover O, Tanne E, Peretz Y, Maori E, Chernin L, Sela I (2012) Expression of an entire bacterial operon in plants. Plant Physiol 158:1883–1892
77. Gaeta RT, Masonbrink RE, Krishnaswamy L, Zhao C, Birchler JA (2012) Synthetic chromosome platforms in plants. Annu Rev Plant Biol 63:307–330
78. Houben A, Mette MF, Teo CH, Lermontova I, Schubert I (2013) Engineered plant minichromosomes. Int J Dev Biol 57:651–657
79. Farré G, Blancquaert D, Capell T, Van Der Straeten D, Christou P, Zhu C (2014) Engineering complex metabolic pathways in plants. Annu Rev Plant Biol 65:187–223
80. Dhar MK, Kaul S, Kour J (2011) Towards the development of better crops by genetic transformation using engineered plant chromosomes. Plant Cell Rep 30:799–806
81. Murata M (2014) Minichromosomes and artificial chromosomes in Arabidopsis. Chromosome Res 22:167–178
82. Birchler JA (2014) Engineered minichromosomes in plants. Curr Opin Plant Biol 19C:76–80
83. Sainsbury F, Benchabane M, Goulet M-C, Michaud D (2012) Multimodal protein constructs for herbivore insect control. Toxins 4:455–475
84. Beaujean A, Ducrocq-Assaf C, Sangwan RS, Lilius G, Bülow L, Sangwan-Norreel BS (2000) Engineering direct fructose production in processed potato tubers by expressing a bifunctional alpha-amylase/glucose isomerase gene complex. Biotechnol Bioeng 70:9–16
85. Jang I-C, Oh S-J, Seo J-S, Choi W-B, Song SI, Kim CH, Kim YS, Seo H-S, Do Choi Y, Nahm BH, Kim J-K (2003) Expression of a bifunctional fusion of the Escherichia coli genes for trehalose-6-phosphate synthase and trehalose-6-phosphate phosphatase in transgenic rice plants increases trehalose accumulation and abiotic stress tolerance without stunting growth. Plant Physiol 131:516–524
86. Kourtz L, Dillon K, Daughtry S, Madison LL, Peoples O, Snell KD (2005) A novel thiolase-reductase gene fusion promotes the production of polyhydroxybutyrate in Arabidopsis. Plant Biotechnol J 3:435–447
87. Tian L, Dixon RA (2006) Engineering isoflavone metabolism with an artificial bifunctional enzyme. Planta 224:496–507
88. Carrington JC, Dougherty WG (1988) A viral cleavage site cassette: identification of amino acid sequences required for tobacco etch virus polyprotein processing. Proc Natl Acad Sci U S A 85:3391–3395
89. Walsh TA, Strickland JA (1993) Proteolysis of the 85-kilodalton crystalline cysteine proteinase inhibitor from potato releases functional cystatin domains. Plant Physiol 103:1227–1234
90. Urwin PE, Levesley A, McPherson MJ, Atkinson HJ (2000) Transgenic resistance to the nematode Rotylenchulus reniformis conferred by Arabidopsis thaliana plants expressing proteinase inhibitors. Mol Breed 6:257–264
91. Urwin PE, McPherson MJ, Atkinson HJ (1998) Enhanced transgenic plant resistance to nematodes by dual proteinase inhibitor constructs. Planta 204:472–479
92. François IEJA, De Bolle MFC, Dwyer G, Goderis IJWM, Woutors PFJ, Verhaert PD, Proost P, Schaaper WMM, Cammue BPA, Broekaert WF (2002) Transgenic expression in Arabidopsis of a polyprotein construct leading to production of two different antimicrobial proteins. Plant Physiol 128:1346–1358
93. Walker JM, Vierstra RD (2007) A ubiquitin-based vector for the co-ordinated synthesis of multiple proteins in plants. Plant Biotechnol J 5:413–421

94. Zhang B, Rapolu M, Huang L, Su WW (2011) Coordinate expression of multiple proteins in plant cells by exploiting endogenous kex2p-like protease activity. Plant Biotechnol J 9:970–981
95. Luke GA, de Felipe P, Cowton VM (2006) Self-processing polyproteins: a strategy for co-expression of multiple proteins in plants. Biotechnol Genet Eng Rev 23:239–252
96. von Bodman SB, Domier LL, Farrand SK (1995) Expression of multiple eukaryotic genes from a single promoter in Nicotiana. Biotechnology (NY) 13:587–591
97. Ceriani MF, Marcos JF, Hopp HE, Beachy RN (1998) Simultaneous accumulation of multiple viral coat proteins from a TEV-NIa based expression vector. Plant Mol Biol 36:239–248
98. Dasgupta S, Collins GB, Hunt AG (1998) Co-ordinated expression of multiple enzymes in different subcellular compartments in plants. Plant J 16:107–116
99. Liang H, Gao H, Maynard CA, Powell WA (2005) Expression of a self-processing, pathogen resistance-enhancing gene construct in Arabidopsis. Biotechnol Lett 27:435–442
100. Saunders K, Sainsbury F, Lomonossoff GP (2009) Efficient generation of cowpea mosaic virus empty virus-like particles by the proteolytic processing of precursors in insect cells and plants. Virology 393:329–337
101. Bedoya L, Martínez F, Rubio L, Daròs J-A (2010) Simultaneous equimolar expression of multiple proteins in plants from a disarmed potyvirus vector. J Biotechnol 150:268–275
102. Wellink J, Verver J, Van Lent J, van Kammen A (1996) Capsid proteins of cowpea mosaic virus transiently expressed in protoplasts form virus-like particles. Virology 224:352–355
103. Luke G (2012) Translating 2A research into practice. Innovations in biotechnology. InTech Open, Rijeka, pp 165–186
104. Geu-Flores F, Olsen CE, Halkier BA (2009) Towards engineering glucosinolates into non-cruciferous plants. Planta 229:261–270
105. van Herpen TWJM, Cankar K, Nogueira M, Bosch D, Bouwmeester HJ, Beekwilder J (2010) Nicotiana benthamiana as a production platform for artemisinin precursors. PLoS ONE 5, e14222
106. François IEJA, Van Hemelrijck W, Aerts AM, Wouters PFJ, Proost P, Broekaert WF, Cammue BPA (2004) Processing in Arabidopsis thaliana of a heterologous polyprotein resulting in differential targeting of the individual plant defensins. Plant Sci 166:113–121
107. Ralley L, Enfissi EMA, Misawa N, Schuch W, Bramley PM, Fraser PD (2004) Metabolic engineering of ketocarotenoid formation in higher plants. Plant J 39:477–486
108. Quilis J, López-García B, Meynard D, Guiderdoni E, San Segundo B (2014) Inducible expression of a fusion gene encoding two proteinase inhibitors leads to insect and pathogen resistance in transgenic rice. Plant Biotechnol J 12:367–377
109. El Amrani A, Barakate A, Askari BM, Li X, Roberts AG, Ryan MD, Halpin C (2004) Coordinate expression and independent subcellular targeting of multiple proteins from a single transgene. Plant Physiol 135:16–24
110. Lee D-S, Lee K-H, Jung S, Jo E-J, Han K-H, Bae H-J (2012) Synergistic effects of 2A-mediated polyproteins on the production of lignocellulose degradation enzymes in tobacco plants. J Exp Bot 63:4797–4810
111. Ma C, Mitra A (2002) Expressing multiple genes in a single open reading frame with the 2A region of foot-and-mouth disease virus as a linker. Mol Breed 9:191–199
112. Sun H, Lang Z, Zhu L, Huang D (2012) Acquiring transgenic tobacco plants with insect resistance and glyphosate tolerance by fusion gene transformation. Plant Cell Rep 31:1877–1887
113. López-Lastra M, Rivas A, Barría MI (2005) Protein synthesis in eukaryotes: the growing biological relevance of cap-independent translation initiation. Biol Res 38:121–146
114. Urwin P, Yi L, Martin H, Atkinson H, Gilmartin PM (2000) Functional characterization of the EMCV IRES in plants. Plant J 24:583–589
115. Urwin PE, Zubko EI, Atkinson HJ (2002) The biotechnological application and limitation of IRES to deliver multiple defence genes to plant pathogens. Physiol Mol Plant Pathol 61:103–108
116. Ha S-H, Liang YS, Jung H, Ahn M-J, Suh S-C, Kweon S-J, Kim D-H, Kim Y-M, Kim J-K (2010) Application of two bicistronic systems involving 2A and IRES sequences to the biosynthesis of carotenoids in rice endosperm. Plant Biotechnol J 8:928–938
117. Gouiaa S, Khoudi H, Leidi EO, Pardo JM, Masmoudi K (2012) Expression of wheat Na(+)/H(+) antiporter TNHXS1 and H(+)- pyrophosphatase TVP1 genes in tobacco from a bicistronic transcriptional unit improves salt tolerance. Plant Mol Biol 79:137–155
118. Peretz Y, Mozes-Koch R, Akad F, Tanne E, Czosnek H, Sela I (2007) A universal expression/silencing vector in plants. Plant Physiol 145:1251–1263
119. Gover O, Peretz Y, Mozes-Koch R, Maori E, Rabinowitch HD, Sela I (2014) Only minimal regions of tomato yellow leaf curl virus (TYLCV) are required for replication, expression and movement. Arch Virol. doi:10.1007/s00705-014-2066-7
120. Oey M, Lohse M, Kreikemeyer B, Bock R (2009) Exhaustion of the chloroplast protein synthesis capacity by massive expression of a highly stable protein antibiotic. Plant J 57:436–445
121. Ruhlman T, Verma D, Samson N, Daniell H (2010) The role of heterologous chloroplast sequence elements in transgene integration and expression. Plant Physiol 152:2088–2104

122. Bock R (2013) Strategies for metabolic pathway engineering with multiple transgenes. Plant Mol Biol 83:21–31
123. Bock R (2007) Plastid biotechnology: prospects for herbicide and insect resistance, metabolic engineering and molecular farming. Curr Opin Biotechnol 18:100–106
124. Jayaraj J, Devlin R, Punja Z (2007) Metabolic engineering of novel ketocarotenoid production in carrot plants. Transgenic Res 17:489–501
125. Verma D, Samson NP, Koya V, Daniell H (2008) A protocol for expression of foreign genes in chloroplasts. Nat Protoc 3:739–758
126. Clarke JL, Daniell H (2011) Plastid biotechnology for crop production: present status and future perspectives. Plant Mol Biol 76:211–220
127. Daniell H (2007) Transgene containment by maternal inheritance effective or elusive? Proc Natl Acad Sci U S A 104:6879–6880
128. Ruf S, Karcher D, Bock R (2007) Determining the transgene containment level provided by chloroplast transformation. Proc Natl Acad Sci U S A 104:6998–7002
129. Caroca R, Howell KA, Hasse C, Ruf S, Bock R (2013) Design of chimeric expression elements that confer high-level gene activity in chromoplasts. Plant J 73:368–379
130. Maliga P, Bock R (2011) Plastid biotechnology food, fuel, and medicine for the 21st century. Plant Physiol 155:1501–1510
131. Scharff LB, Bock R (2014) Synthetic biology in plants. Plant J 78:783–798
132. Kuroda H, Maliga P (2001) Complementarity of the 16S rRNA penultimate stem with sequences downstream of the AUG destabilizes the plastid mRNAs. Nucleic Acids Res 29:970–975
133. Drechsel O, Bock R (2011) Selection of Shine-Dalgarno sequences in plastids. Nucleic Acids Res 39:1427–1438
134. Bock R (2014) Genetic engineering of the chloroplast: novel tools and new applications. Curr Opin Biotechnol 26:7–13
135. Bock R (2014) Engineering chloroplasts for high-level foreign protein expression. In: Maliga P (ed) Chloroplast biotechnology: methods and protocols, vol 1132. Springer, New York, pp 93–106
136. Liere K, Börner T (2007) Transcription and transcriptional regulation in plastids. In: Bock R (ed) Cell and molecular biology of plastids. Springer, Berlin, pp 121–174
137. Maliga P (2004) Plastid transformation in higher plants. Annu Rev Plant Biol 55:289–313
138. Chakrabarti SK, Lutz KA, Lertwiriyawong B, Svab Z, Maliga P (2006) Expression of the cry9Aa2 B.t. gene in tobacco chloroplasts confers resistance to potato tuber moth. Transgenic Res 15:481–488
139. Kuroda H, Maliga P (2001) Sequences downstream of the translation initiation codon are important determinants of translation efficiency in chloroplasts. Plant Physiol 125:430–436
140. Zhang J, Ruf S, Hasse C, Childs L, Scharff LB, Bock R (2012) Identification of cis-elements conferring high levels of gene expression in non-green plastids. Plant J 72:115–128
141. Krichevsky A, Meyers B, Vainstein A, Maliga P, Citovsky V (2010) Autoluminescent plants. PLoS ONE 5, e15461
142. Westhoff P, Herrmann RG (1998) Complex RNA maturation in chloroplasts. Eur J Biochem 171:551–564
143. Hirose T, Sugiura M (1997) Both RNA editing and RNA cleavage are required for translation of tobacco chloroplast ndhD mRNA: a possible regulatory mechanism for the expression of a chloroplast operon consisting of functionally unrelated genes. EMBO J 16:6804–6811
144. Walter M, Piepenburg K, Schöttler MA, Petersen K, Kahlau S, Tiller N, Drechsel O, Weingartner M, Kudla J, Bock R (2010) Knockout of the plastid RNase E leads to defective RNA processing and chloroplast ribosome deficiency. Plant J 64:851–863
145. De Cosa B, Moar W, Lee S-B, Miller M, Daniell H (2001) Overexpression of the Bt cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. Nat Biotechnol 19:71–74
146. Quesada-Vargas T, Ruiz ON, Daniell H (2005) Characterization of heterologous multigene operons in transgenic chloroplasts. Transcription, processing, and translation. Plant Physiol 138:1746–1762
147. Hasunuma T, Miyazawa S-I, Yoshimura S, Shinzaki Y, Tomizawa K-I, Shindo K, Choi S-K, Misawa N, Miyake C (2008) Biosynthesis of astaxanthin in tobacco leaves by transplastomic engineering. Plant J 55:857–868
148. Nakashita H, Arai Y, Shikanai T, Doi Y, Yamaguchi I (2001) Introduction of bacterial metabolism into higher plants by polycistronic transgene expression. Biosci Biotechnol Biochem 65:1688–1691
149. Magee AM, Horvath EM, Kavanagh TA (2004) Pre-screening plastid transgene expression cassettes in Escherichia coli may be unreliable as a predictor of expression levels in chloroplast-transformed plants. Plant Sci 166:1605–1611
150. Zhou F, Karcher D, Bock R (2007) Identification of a plastid intercistronic expression element (IEE) facilitating the expression of stable translatable monocistronic mRNAs from operons. Plant J 52:961–972
151. Lu Y, Rijzaani H, Karcher D, Ruf S, Bock R (2013) Efficient metabolic pathway engineering in transgenic tobacco and tomato plastids with synthetic multigene operons. Proc Natl Acad Sci U S A 110:E623–E632
152. Cerutti H, Osman M, Grandoni P, Jagendorf AT (1992) A homolog of Escherichia coli RecA protein

in plastids of higher plants. Proc Natl Acad Sci U S A 89:8068–8072

153. Krech K, Fu H-Y, Thiele W, Ruf S, Schöttler MA, Bock R (2013) Reverse genetics in complex multigene operons by co-transformation of the plastid genome and its application to the open reading frame previously designated psbN. Plant J 75:1062–1074

154. Kumar S, Hahn FM, Baidoo E, Kahlon TS, Wood DF, McMahan CM, Cornish K, Keasling JK, Daniell H, Whalen MC (2012) Remodeling the isoprenoid pathway in tobacco by expressing the cytoplasmic mevalonate pathway in chloroplasts. Metab Eng 14:19–28

155. Sinagawa-García SR, Tungsuchat-Huang T, Maliga P (2009) Next generation synthetic vectors for transformation of the plastid genome of higher plants. Plant Mol Biol 70:487–498

156. Day A, Goldschmidt-Clermont M (2011) The chloroplast transformation toolbox selectable markers and marker removal. Plant Biotechnol J 9:540–553

18 Transient Expression Systems in Plants: Potentialities and Constraints

Tomas Canto

Abstract

Plants have been used from old to extract and isolate by different means the products of interest that they store. In recent years new techniques have emerged that allow the use of plants as factories to overexpress transiently and often efficiently, specific genes of interest, either endogenous or foreign, in their native form or modified. These techniques allow and facilitate the targeted purification of gene products for research and commercial purposes without resorting to lengthy, time-consuming and sometimes challenging plant stable transformations, while avoiding some of their associated regulatory constraints. In this chapter we describe the main strategies available for the transient expression of gene sequences and their encoded products in plants. We discuss biological issues affecting transient expression, including resistance responses elicited by the plant against sequences that it recognizes naturally as foreign, and ways to neutralize them. We also discuss the relative advantages of each expression strategy as well as their inherent drawbacks and technical limitations, and how to partially prevent or overcome them, whenever possible.

Keywords

Transient expression systems in plants • Transient expression by agroinfiltration • Transient expression in plants by biolistic bombardment • Transient expression from plant virus vectors

18.1 Introduction

Knowledge on compounds of applied interest that some plants produce and store in their tissues, and on procedures developed for their extraction, has been slowly acquired by mankind from ancient times. This knowledge has become

T. Canto (✉)
Centro de Investigaciones Biológicas (CIB, CSIC),
Ramiro de Maeztu 9, 28040 Madrid, Spain
e-mail: tomas.canto@cib.csic.es

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_18

a main component of our cultural heritage. However, the advent some 50 years ago of the molecular biology era, when the molecular structure of DNA fibers was understood [1] and it was discovered that it harbors genetic information, followed by the creation of molecular biology tools for the handling of nucleic acids has revolutionized our approach to obtaining products and traits of interest from organisms, by allowing the targeted manipulation of their genetic expression profiles.

With these new molecular tools plants can be made to theoretically overproduce virtually any product, endogenous or foreign, so long as the plant species is amenable to the manipulation procedures demanded. For a plant to express any gene in such a way, the first step is to introduce it into the plant cells. This could be achieved by stable transformation (see Chap. 17), usually with agrobacterium-delivered T-DNAs, sometimes through bombardment or by other means. Stable plant transformation has however limitations. To mention some, first, established procedures to regenerate transgenic plants from transformed cells in calli derived from plant tissues or from cell cultures are limited to a few plant species. Second, procedures to obtain homozygous transgenic lines may be lengthy, and for example in tomato they could require more than a year. Third, if the product to be expressed is deleterious or harmful to the plant, regeneration of full size, healthy-looking plants may not be possible, or require the use of for example inducible promoters or other specialized approaches. Fourth, licensing the use in the field of transgenic plants carries limitations in those countries/economic areas where they are allowed, as well as lengthy safety and regulatory procedures that would add further years to their actual availability for non-research use.

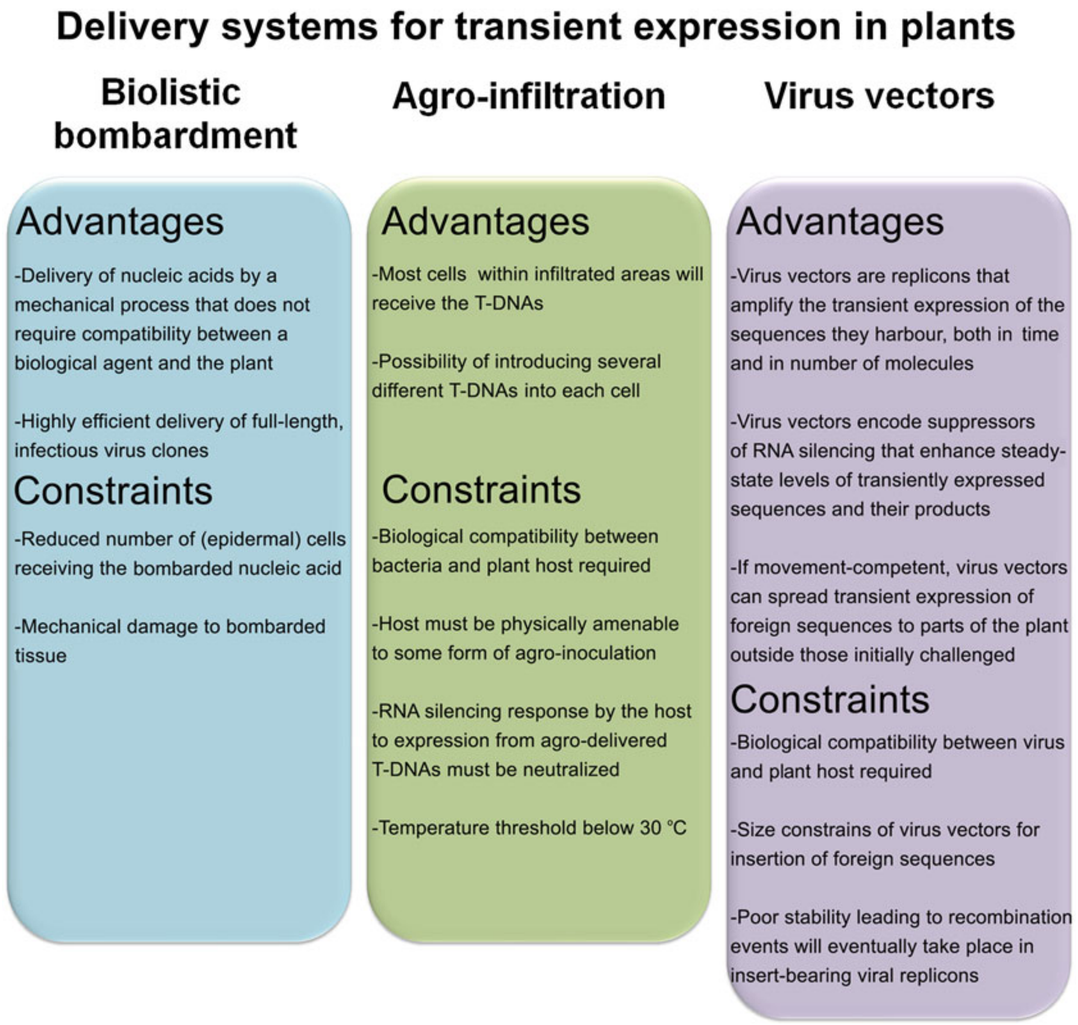
An alternative to plant stable transformation is the use of transient expression systems to express the desired products on already grown, non-transgenic plants. Basically, there are three transient expression delivery systems in plants (Fig. 18.1), plus combinations thereof: (a) the biolistic bombardment of nucleic acids; (b) the agrobacterium-mediated transfer of T-DNA fragments; (c) the use of plant virus vectors. These three major methods for transient expression in plants will be further described in the next sections.

18.2 Biolistic Bombardment

Biolistic bombardment of plant leaves with nucleic acids that encode genes of interest, either as DNAs under the control of plant-compatible eukaryotic promoter and terminator sequences, usually a circular plasmid for convenience and stability, but also linear DNA or PCR products, or alternatively as RNAs, will introduce some of these molecules into epidermal, trichome and even mesophyll cells in live plant leaves. There, they will express transiently the genes they carry.

Bombardment devices deliver the nucleic acids coated on tungsten or gold spherical particles of between 1 and 2 μm in diameter by means of high-pressure shots (commonly around 3 bar). Both shot pressure and metal particles help deliver the nucleic acids into the cells. These particles, however, also cause a degree of mechanical damage to the targeted tissue [2], and only some of the cells where the particles are introduced survive the mechanical stress and express the exogenous genes. The extent of tissue damage and the number of cells that express transiently these genes will depend on parameters such as the type of bombardment gun used, how tender-leaved the plant species is, the distance of the device to the leaf surface, the type of particle used, or the pressure used for shooting. Thus, for every plant species and bombardment device, these parameters of use must be optimized [2–4].

Historically, the origins of the technique date back to the 1980s when Klein et al. (1987) [5] demonstrated at Cornell University that a virus (*Tobacco mosaic virus*, TMV) could be delivered into onion epidermal cells using a laboratory-manufactured Gene Gun bombardment device. Subsequently, other researchers, in particular plant virologists, tested the procedure using different variations of this initial device, either of commercial origin or manufactured by themselves



Delivery systems for transient expression in plants

Biolistic bombardment

Advantages

-Delivery of nucleic acids by a mechanical process that does not require compatibility between a biological agent and the plant

-Highly efficient delivery of full-length, infectious virus clones

Constraints

-Reduced number of (epidermal) cells receiving the bombarded nucleic acid

-Mechanical damage to bombarded tissue

Agro-infiltration

Advantages

-Most cells within infiltrated areas will receive the T-DNAs

-Possibility of introducing several different T-DNAs into each cell

Constraints

-Biological compatibility between bacteria and plant host required

-Host must be physically amenable to some form of agro-inoculation

-RNA silencing response by the host to expression from agro-delivered T-DNAs must be neutralized

-Temperature threshold below 30 °C

Virus vectors

Advantages

-Virus vectors are replicons that amplify the transient expression of the sequences they harbour, both in time and in number of molecules

-Virus vectors encode suppressors of RNA silencing that enhance steady-state levels of transiently expressed sequences and their products

-If movement-competent, virus vectors can spread transient expression of foreign sequences to parts of the plant outside those initially challenged

Constraints

-Biological compatibility between virus and plant host required

-Size constrains of virus vectors for insertion of foreign sequences

-Poor stability leading to recombination events will eventually take place in insert-bearing viral replicons

Fig. 18.1 Transient expression delivery systems in plants: biolistic bombardment; *Agrobacterium tumefaciens*-mediated delivery of T-DNAs; virus vectors (*left*, *central* and *right panels*, respectively), plus combinations thereof, as virus vectors could be delivered into plants either by mechanical rubbing of infectious nucleic acids, by biolistic bombardment, or by agroinfiltration, if expressed from full-length infectious binary constructs. The main potentialities and constraints of each system are indicated in the corresponding panels

in the laboratory, to inoculate into plants full-length infectious viral RNAs (either *in vitro* transcripts or extracted from virus-infected plants) or DNAs (either cDNA clones or true viral DNAs) corresponding to both, RNA or DNA plant viruses.

Traditional means of infection of plants with plant viruses include mechanical rubbing of carborundum- (Silicon carbide, CSi) or celite- (diatomaceous earth, SiO_2) dusted leaves with solutions containing infectious virions or viral nucleic acids. This procedure was and still is widely and successfully used for many viruses, but it is ineffectual in the case of phloem-limited viruses, or when infecting some hard-leaved or woody plants. For these difficult cases, delivery of viruses into plants had been achieved by other means, such as using their specific natural vectors, insects or nematodes, or even by grafting, but these techniques are both, time-consuming and technically demanding. What plant virologists found is that in many cases biolistic bombardment was capable of overcoming these difficulties, as well as an efficient procedure for the delivery and successful infection of different plant species with the viruses tested, in comparison to the traditional means of infection mentioned above [2–4, 6].

While the initial bombardment devices placed the target plant inside a partial vacuum container to facilitate penetration of the particles, newer devices lack these chambers for ease of manipulation, at the expense of somewhat lesser efficiency [4]. Some devices currently under use are available commercially, such as the Bio-Rad Helios® Gene Gun system, while others are manufactured from researchers themselves, such as the HandGun [3], or the HandyGun [2].

18.2.1 Advantages

The main advantage of biolistic bombardment over other transient expression systems is that it delivers nucleic acids into live plant cells through a mechanical process that does not require interaction between the plant species and a compatible biological agent, such as bacteria or viruses. The technique requires adapting bombardment conditions to the specific host plant, to achieve its maximum efficiency.

18.2.2 Constraints

Even though bombardment has proven a more efficient technique than mechanical inoculation to infect plants with some plant viruses, its efficiency as a means to express an introduced gene in as many cells as possible is low. In a bombarded leaf typically only a handful of surviving cells receive and express the foreign nucleic acid, a number that is more in the range of the tens than in the hundreds of cells, as can be seen by the limited fluorescence found by confocal microscopy in *Nicotiana* spp. leaves bombarded with RNAs encoding fluorescent protein markers [7]. Thus, unless the nucleic acids delivered express an infectious agent that can replicate and spread at least locally, and if possible systemically, throughout the plant (*i.e.*, a movement-competent plant virus vector) from the bombarded cell, expression products are constricted to the few initial cells that received the nucleic acid-coated particles, and sometimes to a halo of neighboring cells connected to them by plasmodesmata, which is usually no more than one or two layers thick. This latter effect is likely caused by unrestricted traffic of small proteins expressed at the initial cell through plasmodesmata. Such is the case of free *Aquorea victoria* green fluorescent protein (GFP), of *ca*. 25 kDa, between *N. benthamiana* epidermal cells [7]. In this regard, it appears that proteins up to 50 kDa can traffic freely through simple type plasmodesmata in sink tissues of *Nicotiana* spp. plants before leaf tissue conversion into source alters plasmodesmata types, and drastically reduces their size exclusion limits [8].

Thus, if the bombarded nucleic acid is a plant virus vector that carries the gene of interest, which can spread in that host from the initially bombarded cells and replicate elsewhere, bombardment could be considered an efficient technique to facilitate infection by the virus vector and expression of the gene carried by the vector. By contrast, if the bombarded nucleic acid is non-viral and lacks the ability to replicate and spread into other cells, bombardment should be considered a specialist tool for research applications that study processes at the individual or the cell cluster level, such as microscopy, and where other options are either not possible or advisable; for example, because the presence of a biological agent (a virus, a bacteria) interferes with the purpose of the research. Otherwise, for research or biotechnology applications that would require large amounts of plant tissue expressing the foreign gene, bombardment would likely be too inefficient and the two other means of transient expression through biological agents would be preferable.

18.3 Agrobacterium-Mediated Transfer of T-DNA Fragments into Plant Cells

The introduction of DNA fragments through agrobacterium-mediated transfer of T-DNA was found to be a powerful research tool that allows the transient expression of any gene in a plant [9–12] after *Agrobacterium tumefaciens* had become of routine use to transform plants stably and constitutively. *A. tumefaciens* is one of the few bacteria capable of delivering DNAs (transfer DNAs, or T-DNAs) into plants. T-DNA delivery involves a complex set of bacterial genes, and the formation of a physical pilus structure that allows the transfer of bacterial DNA into plant cells. In nature, the T-DNA fragment of the tumor-inducing (Ti) plasmids transferred from the agrobacterium into plant nuclei encodes genes required for crown gall tumor formation as part of the bacterial life cycle. However, laboratory modifications have created the numerous versatile binary vector systems currently available, which are composed of pairs of plasmids: the helper plasmid incorporated in the agrobacterium strain that carries many of the Ti plasmid essential genes that allow T-DNA transfer into plant cells, and the binary vector, that carries the T-DNA, free of tumor-inducing genes. Instead, binary vectors can now carry any desired gene or sequence fragment under the control of a eukaryotic promoter, the most common of which is for research purposes the *Cauliflower mosaic virus* 35S promoter, plus a terminator sequence. The binary vector is compatible with both, agrobacterium and *Escherichia coli* and therefore can be manipulated and modified in the latter host like any other *E. coli* plasmid, by standard molecular cloning techniques [13].

As mentioned, the T-DNA molecules delivered into a plant cell that integrate stably by recombination into the plant nuclear genome become inheritable and their selection constitutes the basis of the most commonly used technique for plant transformation [14]. On the other hand, transient expression of genes and their products from binary T-DNAs in leaf tissue infiltrated with the agrobacterium culture constitutes a technique commonly known as agroinfiltration or agroinjection. Cultures could also be inoculated with a needle or stick (agroinoculation). Expression of reporter genes and of non-coding sequences in the infiltrated tissue (the agropatch) from T-DNA fragments has been studied in some detail [15, 16] and ways to enhance their levels of expression or its large-scale use have been envisaged, using a variety of approaches [17, 18].

The agrobacterium infiltration procedure involves the exponential growth of the bacterial culture at 28 °C, from either frozen stock or from individual plate colonies, and its scaling up to the desired final volume until it reaches an Absorbance or Optical Density (OD) at 600 nm of between 1 and 2. Above 30 °C the bacteria in the culture loses the binary vector, thus becoming a limiting threshold for culture growth. Growth is achieved in the selective presence of at least the antibiotic for which resistance is conferred by the binary vector that harbors the T-DNA, although additional antibiotic resistance from the helper plasmid or even chromosomal resistance may also be added. Cultures are pelleted and resuspended in a solution that contains acetosyringone (4′-Hydroxy-3′,5′-dimethoxyacetophenone), which will induce the expression of bacterial genes that will facilitate the T-DNA transfer process [14]. Exposure to acetosyringone typically lasts between 2 and 3 h. Cultures are diluted to the desired OD and infiltrated into plant leaves using a needleless syringe. In most cases, typical infiltration ODs range between 0.2 and 0.5 for optimal product expression [19–21], although in some works ODs as high as 2 (particularly in earlier works) or below 0.1 have been used. By personal experience no apparent differences in protein expression were found using culture ODs between 2 and 0.2, suggesting that in the former, a large excess of bacteria was being unnecessarily infiltrated. Bacterial cultures carrying different binary constructs that express different products can be mixed and co-infiltrated together at the same or at different ODs to guarantee co-expression of different genes in the same cells [15–17, 22]. Co-infiltrations of two or three cultures are common practice in plant pathology and plant biology research, and allow the study of protein-protein interactions, protein co-localizations, the use of one of the expressed products as marker to specific subcellular structures, or as suppressor of defensive responses of the plant to T-DNA expression (see below).

In contrast to infiltration ODs, transient expression levels display a curve of accumulation that may be different for each product expressed from an agro-delivered T-DNA. In most cases, maximum levels of expression occur at 3–4 days after infiltration and fade rapidly after 5–6 days, but this must be confirmed empirically for each gene product. Expression levels will depend on factors such as the strength of the silencing resistance response of the host plant to the particular T-DNA sequence that will affect steady-state levels of transcript T-DNA-derived messenger RNA (mRNA) levels (see below), and also on the intrinsic stability and turnover of the protein product in the cellular environment, whether it is degraded by routes such as the proteasome or autophagy. Protein accumulation in the first 24 h after infiltration (hpi) is usually low and often undetectable [23] but this is not necessarily always the case. In fact for some proteins, the maximum accumulation has been described as early as 24 hpi, possibly for any of the reasons mentioned above [22]. There seems to be no direct relationship between size of the protein product and the time it takes to accumulate and reach its peak after infiltration [22]. Thus a time-course accumulation analysis is advisable for each new protein product being expressed.

18.3.1 Advantages

The main advantage of agroinfiltration over biolistic bombardment is that most plant cells inside the area infiltrated with the bacterial culture will receive the T-DNAs and express the desired genes. This allows the simple scale-up of the procedure by increasing the infiltrated surfaces [18] to produce large amounts of the T-DNA-derived product/s, thus opening the possibility of large-scale applications. In addition, agroinfiltration provides the possibility of expressing more than one product in the majority of the cells in the infiltrated patches, by using mixtures of bacterial cultures harboring different binary constructs. This is problematic using bombardment, or from virus vectors because of cross-protection preventing similar viruses from being simultaneously within the same cell, unless all different products are expressed from the same virus.

18.3.2 Constraints

Transient, steady-state levels of gene products expressed from T-DNAs delivered into the infiltrated leaf patch (agropatch) are influenced by several factors: Choice of plant host is an important one. While choosing the host may not be possible to research performed on a particular plant species, for biotechnology applications in which transient levels of the genes produced and the ease to isolate them are the main issue, careful selection of host is important. Some plant species are not amenable to physical infiltration of their leaves with agrobacterium cultures or may not be compatible with the bacteria. The experimental plant species *Arabidopsis thaliana*, *Nicotiana tabacum*, or *Nicotiana benthamiana* can all three be infiltrated by the means of syringing, but differences in the respective transient levels of the gene products achieved are rather large: *N. benthamiana* expresses higher levels of transcript mRNAs and their products than the other two ones [24]. This could be related to its having naturally truncated the salicylic acid, virus-inducible RNA-dependent RNA polymerase 1 (RdRP1) involved in antiviral defenses, perhaps causing its hypersusceptibility to many different plant viruses [25]. RdRP1 in tobacco on the other hand has been shown to have suppression of silencing activity [26]. Thus, unless a study requires a specific plant species, *N. benthamiana* is a good host of choice for agroinfiltration assays in both, experimental and biotechnology studies [18, 27].

Another important factor that constrains transient expression from T-DNAs is their being recognized as foreign by the plant, which elicits an RNA-based silencing response that depresses both, the steady-state levels of the transcript messenger RNA (mRNA) encoded by the T-DNA and those of the protein product it may encode [15]. The trigger of this silencing resistance is most likely the presence of double-stranded transcripts derived from sense and antisense transcription of the T-DNA sequences, causing the generation by the RNA silencing machinery of the plant of small interfering RNAs to even promoter sequences, or to promoter-less T-DNAs, in theory not expected to be transcribed [16]. These small RNAs will guide host protein complexes to which they bind towards RNAs with whom they have sequence complementarity, resulting in the slicing and destruction of the latter [28]. To neutralize this silencing response and enhance the transient steady-state levels of T-DNA encoded genes, co-expression of proteins that are capable to interfere with components of this resistance is used routinely. These factors are known as "suppressors" of RNA silencing. Most if not all plant viruses express at least one suppressor factor, as for several reasons all DNA or RNA plant viruses induce during their life cycle dsRNAs that trigger a plant RNA-based antiviral silencing response. Left unchecked, silencing would have devastating consequences to the virus and provide the plant with immunity to infection. Viral suppressor factors were discovered in the late 1990s of the past century [29] soon after the small RNA-based defense and regulation system involved not only in biotic resistance, but also in plant development and in responses to the environment, was itself discovered. To the date more than 35 viral proteins have been identified as suppressors of silencing. Use of viral suppressors of gene silencing to prevent the targeted degradation of infiltrated T-DNA-derived transcripts by gene silencing was empirically shown to counteract this gene regulatory and resistance system [15–17] and is now routinely used to that purpose (Fig. 18.2).

One of the main ways to determine the strength of the suppression of silencing of a viral suppressor is by expressing it from T-DNAs together with a reporter gene, such as GFP expressed from a separate T-DNA, and checking the steady-state levels of reporter achieved either in the presence or in the absence of the suppressor. This biological assay is called by plant virologists "agropatch suppressor assay". Depending on how much suppressors prevent the partial silencing of the reporter they have been characterized as weak, such are the *Potato virus X* (PVX) p25 protein [30], or the tobravirus *Tobacco rattle virus* (TRV) 16K protein [31], or as strong, such as most potyviral HCPros, *Tomato bushy stunt virus* (TBSV) P19 [32] or the 2b protein from some *Cucumber mosaic virus* (CMV) strains, for example [19, 23]. To achieve maximum transient expression from agroinfiltrated patches, the use of a strong suppressor of silencing would in principle be advisable. However, if this expression was to be achieved from an agro-delivered virus vector rather than from T-DNAs that are not movement-competent replicons, then this would not have to be necessarily the case, as will be seen in the next section. Regarding use of suppressors to enhance transient expression levels from agrodelivered T-DNAs, it should also be noted that in the evolutionary race between plants and viruses, some plant species have evolved extreme resistances to specific viruses triggered by their small RNA-binding suppressors [30] that in some circumstances should be considered, if one encounters an immunity or necrotic response to infiltration with a particular suppressor.

As agrodelivered T-DNAs trigger a silencing response to any genetic sequence that is present in the T-DNA, it should be noted that any sequences in the plant that share sequence similarity with them, either endogenous genes, or terminator sequences in stably-transformed transgenes, will also be targeted for silencing in the infiltrated patches [28]. This fact allows the targeted, transient and partial silencing of plant genes in infiltrated tissues. This silencing will often go to the whole plant in the case of integrated transgenes, but not so in those of endogenous genes, for reasons not well understood. If silencing of endogenous genes were the aim of infiltration, then co-expression of a suppressor would be naturally not advisable, as it would reduce or prevent the silencing response.

Temperature is a third factor that constrains agroinfiltration as a tool for gene expression in

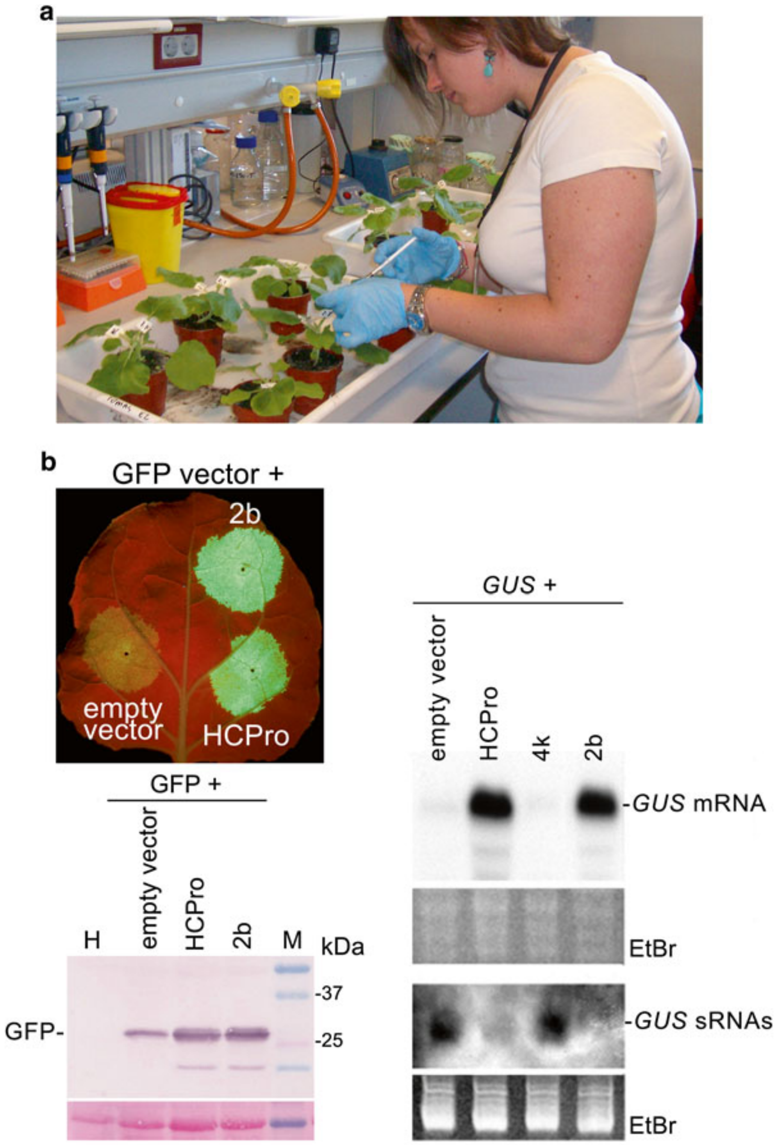


Fig. 18.2 Transient expression in plants by agroinfiltration. (**a**) Circular patches in young, fully expanded leaves of *Nicotiana benthamiana* leaves become infiltrated with agrobacterium cultures harboring binary vectors using a needleless syringe. (**b**) Transient expression by agroinfiltrated patches of two reporters (*Aquorea victoria green* fluorescent protein, GFP and *Escherichia coli* β-glucuronidase, GUS) either in the presence or in the absence of viral suppressors of RNA silencing (HCPro from the *Potyvirus Potato virus Y* and 2b protein from the *Cucumovirus Cucumber mosaic virus*) at 3 days post infiltration. GFP-derived fluorescence could be detected under the UV lamp in infiltrated patches of intact leaves (leaf panel). In the patch infiltrated with a bacterial culture harboring the GFP binary construct mixed with a culture harboring an empty binary construct GFP-derived fluorescence and steady-state levels of GFP were much lower than in those patches co-infiltrated with cultures harboring binaries expressing either HCPro or 2b protein suppressors of RNA silencing (left patch vs. right patches in leaf panel, and corresponding bands in the *left* western blot panels below). Panels below the western blot show Ponceau S-stained membranes after blotting, as controls of loading. Similarly, in patches infiltrated with a binary construct expressing GUS, steady-state levels of *GUS* mRNAs were higher when co-infiltrated with binary constructs expressing HCPro or 2b protein (*upper right* northern blot panel), while the RNA silencing-induced small RNA levels to *GUS* sequences were reduced in the presence of the viral suppressors (*lower right* northern blot panel). Ethidium bromide (EtBr) stained gels appear as loading controls. Keys to other symbols: *H* non-infiltrated plant sample, *M* protein molecular weight markers, *4k* a potexviral protein without suppressor of silencing function

plants. Optimal temperatures for transient gene expression through agroinfiltration appear to be in the 25 ± 0.5 °C range in *N. benthamiana* [18, 33]. Temperatures of 29 °C and above prevent development of tumors caused by the agrobacterium as certain proteins involved in the transfer machine are not functional and critically, pilus formation does not take place either [14, 34]. Further to this, it is known that the strength of the plant RNA-based silencing defense against both viruses and T-DNA transcripts increases with temperature [35–38]. Thus, in addition to reduced T-DNA transfer process, stronger silencing responses at higher temperature would negatively affect any expression from agrodelivered T-DNAs. Therefore, agroinfiltration as a technique to transiently express genes in plants at temperatures above 29 °C would appear as a nonviable option. Recently, however, a procedure has been developed that allows transient gene expression in plants from agroinfiltrated T-DNAs at temperatures above that threshold, by providing a 24 h window after infiltration to allow for the T-DNA to be transferred to the plant [23].

18.4 Use of Plant Viruses as Expression Vectors

Many viral vectors have been generated from plant viruses and this section cannot attempt to present them all. Instead, it aims at describing their generic properties, limitations and advantages as expression vectors. There are many types of plant viruses: some have genomic RNAs, others are DNA-based, and both can be either single- or double-stranded. Most plant viruses encapsidate as either isometric virions, or as helical rod- or filament-shaped virions. A few uncommon ones, such as vasculature-confined members of the genus *Umbravirus*, do not even have coat proteins nor do they form virions on their own; instead, they use coat proteins from "assistor viruses" to produce virions. Some plant viruses have a single encapsidating genomic nucleic acid, others have multipartite genomes. Some infect systemically the majority of the host tissues, while others are limited to specific tissues, such as the vasculature. And finally, some have the ability to infect hundreds of plant species from different families, while others have a very restricted host range [Association of Applied Biologists (aab) description of plant viruses: http://www.dpvweb.net/dpv; 39].

Strategies for gene expression in plant viruses are also diverse. Some viruses express their different gene products from individual subgenomic RNAs, such as for example CMV [40], or using internal translation initiation sites within the same RNA, while on the other extreme *Potyviruses* encode all but one of its products as a single gene that expresses a large polyprotein that will undergo post-translational proteolytic processing to generate the ten different final proteins [41].

Despite their diversity, most plant viruses share a remarkable feature that differentiates them from many animal viruses: they are compact and small-sized. Most plant virus genomes fall within the ranges of 3–7 kb in length and the largest of them, those within the genus *Closterovirus* are ~20 kb in length. Consequences of such compactness are: (1) that in many viruses, genes overlap in the same nucleic acid stretch in different reading frames or transcription reading senses; and (2) that many plant viral proteins are multifunctional and important in more than one way to the virus infectious cycle. These facts are of relevance to the development of virus vectors to express foreign sequences, as they will impose limits to their capabilities to act both as fully functional viruses and as expression vectors.

Plant virus vectors were developed from full-length infectious clones of plant viruses after they were first obtained. Historically, the origins of infectious clones of plant RNA viruses date back to the mid-1980s and early 1990s of the past century. At that time, cDNAs from complete viral genomes were cloned into plasmids under the control of bacteriophage promoters (T7, T3, SP6 RNA polymerases), which could be used to generate *in vitro* viral RNA transcripts. With the proper modifications (such as 5′-end capping or polyadenine tails, depending on the virus) those transcripts would become infectious when inoculated into plants [40, 42, 43]. Later on, many of those full-length clones would be transferred into plasmids under the control of eukaryotic promoters to directly inoculate plants with them, avoiding the *in vitro* transcript step, which is time consuming and costly, as single-stranded transcript RNAs are susceptible to degradation by RNases in the environment, and as the processivity of these polymerases is not outstanding, making it difficult to obtain good yields of longer transcripts. These eukaryotic promoter-dependent, full-length infectious clones would be delivered into the plant cells either by biolistic bombardment or by agroinfiltration.

Most full-length infectious virus clones thus generated have been modified and tested as expression vectors, partly because of the insertion of tracking reporters for research purposes. However, limitations in most of them have resulted in only a few of them being routinely used for the expression of foreign genes, or alternatively for the silencing of endogenous genes (virus-induced gene silencing; VIGS) in plants.

18.4.1 Advantages

Advantages of plant virus vectors over other transient expression systems lay on the fact that they are replicons that within the plant cell multiply their copies and greatly amplify the steady-state levels of any foreign gene they may carry, in comparison to those achieved by for example, a non-replicating T-DNA. In addition, as plant viruses encode suppressor of silencing factors, they depress the silencing response of the plant, further increasing gene expression levels. Plant RNA virus replicons expressed from binary constructs can also be modified for optimal expression in all the cells in the infiltrated patch, boosting thus production [21]. If in addition to this, the viral vector remains competent for local movement or even for systemic movement, then expression can also be achieved in plant tissues outside the area initially challenged.

18.4.2 Constraints

To create any virus vector that expresses a foreign gene, manipulations of viral genomes need to take into account two issues: the specific translational strategy of the virus, and the size limitations imposed on their genomes by encapsidation into virion particles. For example, with regard to translational strategy, in the case of *Potyviruses*, which as mentioned express a single polyprotein, insertion of an additional product also requires the addition of flanking motifs that will be recognized by the viral proteases that slice the products of the polyprotein. In contrast to viruses that express their genes from subgenomic RNAs, such as *Potexviruses*, in *Potyviruses* insertion of any additional gene requires also that of a promoter sequence. Alternatively to expressing the foreign gene separately, the protein of interest could also be expressed as a fusion to either terminus of a non-structural viral gene, or more frequently to the viral CP. In this latter case, fusions to the CP of small peptide sequences have been expressed successfully in several virus vectors in what has been called epitope presentation [44].

Limitations to the size of the genomes that can be encapsidated into virions must also be taken into account when inserting a foreign gene as in most cases inability to encapsidate impairs virus local and systemic movement in plants. This is particularly true for isometric virions in both, DNA or RNA viruses, which impose strict limitations to the size of the genomic nucleic acids that can be encapsidated. However, this is not always the case, as size constrains do not prevent the isometric *Apple latent spherical virus* from being an efficient vector for VIGS [45]. An example of an isometric RNA virus is CMV, in which insertion of a *GFP* reporter in one of its three encapsidating RNAs, either as an additional gene or replacing its *movement protein* (*MP*) or its *coat protein* (*CP*) genes led to the virus not being able to spread locally and systemically throughout the plant [7]. Vectors based on isometric DNA begomoviruses, such as those based on *Bean yellow dwarf virus* are used to express desired genes, but at the cost of removing the viral *MP* and *CP* genes required for its spread [27]. Nevertheless, in combination with agroinfiltration these isometric vectors can be used as local replicons that can potently amplify the transient, steady-state levels of expression of the desired gene within the infiltrated patch [27]. An alternative to these size constrains is the removal of viral genes to provide space to the foreign insert and their functional complementation in trans from a stably-integrated transgene expressed in the plant [46].

Genome size constraints are not as strict for rod- or filament-shaped viruses, as they can elongate their virions to accommodate the inserted sequence. It is therefore not surprising that the most frequently used, movement-competent viral vectors are based on messenger-type RNA viruses that display helical packaging, either rod- or filament-shaped virions (Fig. 18.3). These include members of the *Potex*-, *Poty*- or *Tobamovirus* genera, as well *Tobraviruses*. Choice of the vector will depend on whether virus and host are compatible, and also in compatible interactions on the trade-off between severity of infection symptoms induced vs. the virus titer achieved and consequent expression of the foreign sequence.

The *Potexvirus* type member PVX causes infection symptoms that in *Nicotiana* spp. are usually milder than those induced by *Potyviruses* or *Tobamoviruses*. PVX expresses a p25 suppressor of silencing considered as "weak" [30]. PVX vectors were created by adding a new subgenomic RNA with a multiple cloning site in the corresponding cDNA clone, downstream a duplicated promoter sequence obtained from another *Potexvirus* member, to prevent early removal of the added gene by homologous recombination [47, 48]. Alternatively, a GFP reporter was also expressed as a fusion to the viral CP, linked through the *Foot-and-mouth disease virus* 2A catalytic peptide, giving rise to virions that were partially decorated with GFP-CP fusions, as well as to free GFP [49]. Similar results were obtained on a vector based on the *Potexvirus Pepino mosaic virus* [50].

Other RNA viruses such as TMV (43) are also successfully used as vectors. Like potexvirus vectors, TMV vectors follow the strategy of the duplicated promoter and have been successfully used for large-scale expression and analysis of protein libraries, or reporters [25, 51]. Optimization of TMV vectors to achieve full infection of all cells in infiltrated tissues and optimal reporter expression (magnifection) has been set up in *Nicotiana* spp. [25]. Vectors based on the filamentous *Potyviruses* have also been developed using strategies that insert foreign gene between two products in the polyprotein gene sequence, with flanking motifs recognized by the viral proteases that process it post-translationally [52]. In some cases, by inserting multicassette cloning sites, expression of multiple genes in the same cell from a single vector can be achieved [46]. This is an interesting approach, as it is known that in plants infected with two viral vectors that differ only in the insert they carry most cells will multiply either one or the other viral genome, while the number of cells where there is co-infection of both constructs is limited, and reduces progressively as colonization progresses [53].

The *Tobravirus* type member, *Tobacco rattle virus* (TRV), has a bipartite genome that encapsidates as two separate rod-shaped virions. RNA 1 contains replication genes and the viral *MP*, and can replicate and spread within a compatible host independently from RNA 2. RNA 2 contains the viral *CP* gene, plus genes required for the horizontal transmission of viruses between plants by nematode vectors. Uncommon to plant viruses, these latter genes do not seem to play any additional role in the virus cycle within the host, and can thus be removed and replaced by foreign genes, such as GFP at the expense of losing its vector transmissibility between hosts [54]. An interesting feature of TRV is that in *Nicotiana* spp. it causes very mild infection symptoms [55]. The reason for this effect may lay in the fact that it expresses a weak suppressor of silencing [31] that cannot efficiently suppress the antiviral silencing response of the plant. The consequence is that virus levels (and symptoms) become depressed but not suppressed after initial infection, entering a plant "recovery" phase where the virus is still able to spread to most parts of the plant at low levels, expressing its products without inducing the strong infection symptoms caused by other viruses, such as stunting, leaf distortion, chlorosis or even necrosis. For these reasons, this is the vector of choice when silencing by VIGS endogenous plant genes [55].

Although viruses with helical structures allow the insertion of foreign genes, this comes at the

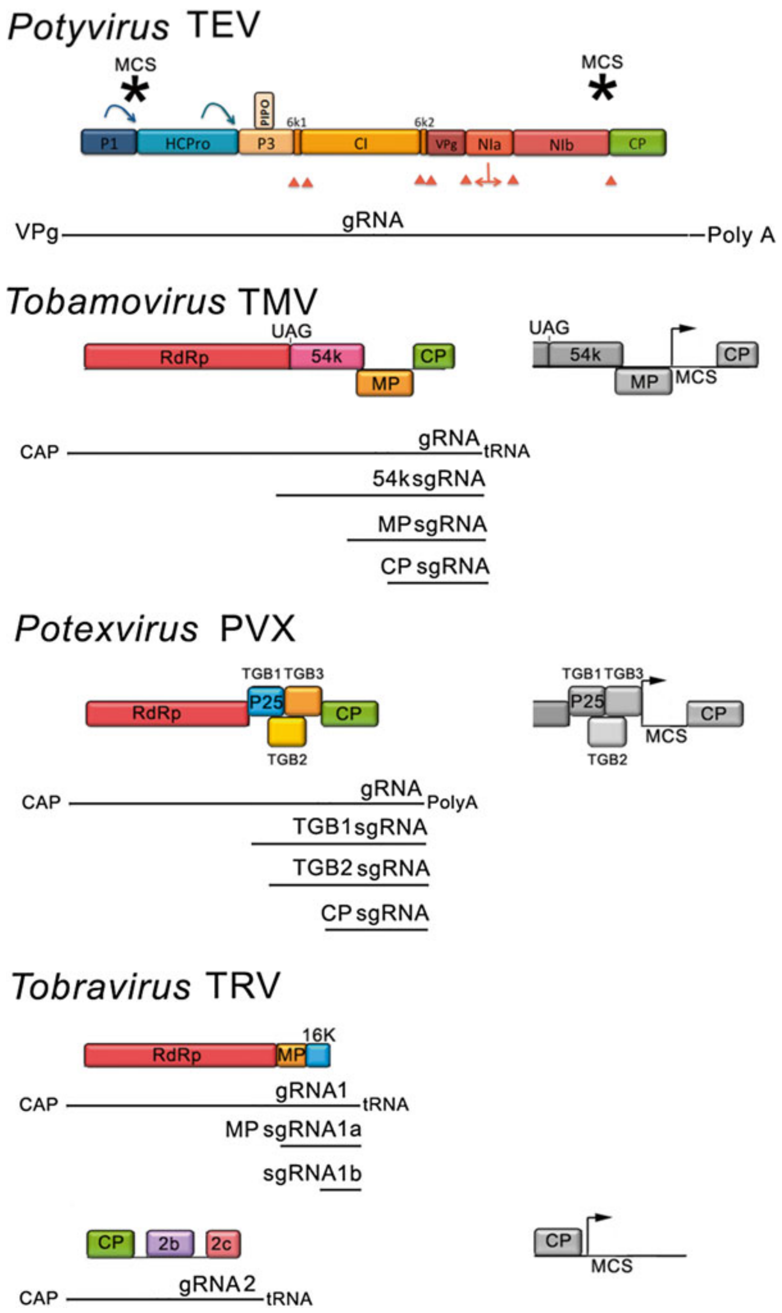


Fig. 18.3 Schematic representation of some of the viral vectors most frequently used for transient expression in plants. They belong to the genera *Potyvirus*, *Tobamovirus*, *Potexvirus* and *Tobravirus*, and are all positive-sense, messenger-type RNA viruses, of helical encapsidation structure, and movement-competent in compatible hosts. In potyvirus vectors, foreign sequences are inserted within the single polyprotein gene, flanked by recognition motifs of viral proteases. Three viral proteases intervene in the post-translational processing of the viral polyprotein: P1 and HCPro cleave themselves at their C termini, while NIa cleaves in cis- and trans- the remaining sites, indicated by spikes. Asterisks indicate two of the most common sites of insertion of foreign sequences. In tobamo-, potex- and tobravirus vectors, expression of foreign sequences is commonly achieved from an inserted duplicated *coat protein* (CP) promoter (indicated by an arrow) from a different species within the genus, to avoid homologous recombination, followed by a multiple cloning site (MCS) for gene insertion, and its expression from a new

price of slower virus movement and virus titers. This may be caused by slower replication, higher exposure of viral RNA to antiviral silencing, slower cell-to-cell movement or loading-unloading into-from the vasculature for systemic movement. As a general rule, the larger the insert, the bigger the detrimental effect observed. In addition to this, recombination events in RNA viruses tend to eject over time the foreign inserts to restore viral fitness. This was common on early vectors, but they improved in their stability by making use of divergent nucleotide sequences when adding additional subgenomic promoters or new protease recognition motifs, in order to prevent homologous recombination events. Even with these precautions, it is a matter of when rather than if recombination and insert removal takes place.

18.5 Conclusions

Transient expression systems in plants have been developed and improved during the last years, to become powerful tools for the expression of different types of products. We have overviewed the main approaches of the delivery and expression of foreign nucleic acid sequences in plants, their evolution and their properties. These approaches have been used in both research and applied contexts to express very large amount of specific products, some with pharmacological and medical applications that are beyond the scope of this chapter to describe. These products take advantage of the relative similarity of post-translational maturation between plants and mammalians in comparison with expression from bacteria. These products include a multitude of modified and recombinant proteins that can be isolated by affinity binding through their tagged epitopes for both research and commercial purposes, the expression of viral particles decorated on their surface with peptides for vaccine production or other purposes [56], antibody production (plantibodies [57]), or the modification (silencing/activation) of metabolic routes in the plant by the targeted silencing of endogenous genes by VIGS routinely used by plant pathologists and plant biologist to study molecular pathways in plants.

Acknowledgements TC's research is currently supported by grant PJ009461 from the Rural Development Administration (RDA) of the Republic of Korea, in cooperation with the Spanish Council for Scientific Research (CSIC), and also by grant BIO2013-47940-R from the Spanish Ministry of Economy and Competitiveness (MEC).

References

1. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737–738
2. Sikorskaite S, Vuorinen AL, Rajamäki M-L, Nieminen A, Gaba V, Valkonen JPT (2010) HandyGun: an improved custom-designed, non-vacuum gene gun suitable for virus inoculation. J Virol Methods 165:320–324
3. Gal-On A, Meiri E, Huet H, Hua WJ, Raccah B, Gaba V (1995) Particle bombardment drastically increases the infectivity of cloned cDNA of zucchini yellow mosaic potyvirus. J Gen Virol 76:3223–3227
4. Gaba V, Lapidot M, Gal-On A (2013) HandGun-mediated inoculation of plants with viral pathogens for mechanistic studies. In: Sudowe S, Reske-Kunz AB (eds) Biolistic DNA delivery: methods and protocols, methods in molecular biology, vol. 940, doi:10.1007/978-1-62703-110-3_5, © Springer Science+Business Media, LLC 2013
5. Klein TM, Wolf ED, Wu R, Sanford JC (1987) High-velocity microprojectiles for delivering nucleic acids into living cells. Nature 327:70–73
6. Gal-On A, Meiri E, Elman C, Gray DJ, Gaba V (1997) Simple handheld devices for the efficient infection of

Fig. 18.3 (continued) subgenomic RNA (*grey* schemes to the *right*). In the case of tobravirus vectors, the inserted promoter and MCS also replace viral genes involved in horizontal transmission by nematode vectors. Poty-, potex- and tobamovirus vectors are used for expression of proteins, whereas the tobravirus vector is more used to silence host sequences (viral-induced gene silencing, VIGS). Key to other symbols: *TEV* Tobacco etch virus, *TMV* Tobacco mosaic virus, *PVX* Potato virus X, *TRV* Tobacco rattle virus, *gRNA, sgRNA* genomic and subgenomic RNAs, respectively, *RdRP* viral RNA-dependent RNA polymerase, *MP* viral movement protein, *HCPro* 16K and P25 are viral suppressors of RNA silencing. Not to scale

plants with viral encoding constructs by particle bombardment. J Virol Methods 64:103–110

7. Canto T, Prior D, Hellwald K, Oparka K, Palukaitis P (1997) Characterization of Cucumber mosaic virus IV. Movement protein and coat protein are both essential for cell-to-cell movement of cucumber mosaic virus. Virology 237:237–248
8. Oparka KJ, Roberts AG, Boevink P, Santa Cruz S, Roberts I, Pradel KS, Imlau A, Kotlizky G, Sauer N, Epel B (1999) Simple, but not branched, plasmodesmata allow the nonspecific trafficking of proteins in developing tobacco leaves. Cell 97:743–754
9. Scofield SR, Tobias CM, Rathjen JP, Chang JH, Lavelle DT, Michelmore RW, Staskawicz BJ (1996) Molecular basis of gene-for-gene specificity in bacterial speck disease of tomato. Science 274:2063–2065
10. Tang XY, Frederick RD, Zhou JM, Halterman DA, Jia YL, Martin GB (1996) Initiation of plant disease resistance by physical interaction of AvrPto and Pto kinase. Science 274:2060–2063
11. Bendahmane A, Kanyuka K, Baulcombe DC (1999) The *Rx* gene from potato controls separate virus resistance and cell death responses. Plant Cell 11:781–791
12. Bendahmane A, Querci M, Kanyuka K, Baulcombe DC (2000) *Agrobacterium* transient expression system as a tool for the isolation of disease resistance genes: application to the *Rx2* locus in potato. Plant J 21:73–81
13. Komori T, Imayama T, Kato N, Ishida Y, Ueki J, Komari T (2007) Current status of binary vectors and superbinary vectors. Plant Physiol 145:1155–1160
14. Gelvin S (2003) *Agrobacterium*-mediated plant transformation: the biology behind "gene-jockeying" tool. Microbiol Mol Biol Rev 2003:16–37
15. Johansen LK, Carrington JC (2001) Silencing on the spot. Induction and suppression of RNA silencing in the *Agrobacterium*-mediated transient expression system. Plant Physiol 126:930–938
16. Canto T, Cillo F, Palukaitis P (2002) Generation of siRNAs by T-DNA sequences does not require active transcription nor homology to sequences in the plant. Mol Plant Microbe Interact 15:1137–1146
17. Voinnet O, Rivas S, Mestre P, Baulcombe D (2003) An enhanced transient expression system in plants based on suppression of gene silencing by the P19 protein of tomato bushy stunt virus. Plant J 33:949–956
18. Chen Q, Lai H, Hurtado J, Stahnke J, Leuzinger K, Dent M (2013) Agroinfiltration as an effective and scalable strategy of gene delivery for production of pharmaceutical proteins. Adv Tech Biol Med 1:103. doi:10.4172/atbm.1000103
19. González I, Martínez L, Rakitina DV, Lewsey MG, Atencio FA, Llave C, Kalinina NO, Carr JP, Palukaitis P, Canto T (2010) Cucumber mosaic virus 2b protein subcellular targets and Interactions: their significance to RNA silencing suppressor activity. Mol Plant-Microbe Interact 23:294–303
20. Bedoya L, Martínez F, Orzáez D, Darós JA (2012) Visual tracking of plant virus infection and movement using a reporter MYB transcription factor that activates anthocyanin biosynthesis. Plant Physiol 158:1130–1138
21. Marillonnet S, Thoeringer C, Kandzia R, Klimyuk V, Gleba V (2005) Systemic Agrobacterium tumefaciens-mediated transfection of viral replicons for efficient transient expression in plants. Nat Biotechnol 23:718–723
22. Liu L, Zhang Y, Tang S, Zhao Q, Zhang Z, Zhang H, Dong L, Guo H, Xie Q (2010) An efficient system to detect protein ubiquitination by agroinfiltration in *Nicotiana benthamiana*. Plant J 61:893–903
23. Del Toro F, Tenllado F, Chung B-N, Canto T (2014) A procedure for the transient expression of genes in plants by agroinfiltration above the permissive threshold to study temperature-sensitive processes in plant-pathogen interactions. Mol Plant Pathol 15:848–857
24. Andrews LB, Curtis WR (2005) Comparison of transient protein expression in tobacco leaves and plant suspension culture. Biotechnol Prog 21:946–952
25. Yang S-J, Carter SA, Cole AB, Cheng N-H, Nelson RS (2004) A natural variant of a host RNA-dependent RNA polymerase is associated with increased susceptibility to viruses by *Nicotiana benthamiana*. Proc Natl Acad Sci U S A 101:6297–6302
26. Ying X-B, Dong L, Zhu H, Duan C-H, Du Q-S, Lv D-Q, Fang Y-Y, García JA, Fang R-X, Guo H-S (2010) RNA-dependent polymerase 1 from *Nicotiana tabacum* suppresses RNA silencing and enhances viral infection in *Nicotiana benthamiana*. Plant Cell 22:1358–1372
27. Chen Q, He J, Phoolcharoen W, Mason HS (2011) Geminiviral vectors based on bean yellow dwarf virus for production of vaccine antigens and monoclonal antibodies in plants. Hum Vaccines 7:331–338
28. Ruiz-Ferrer V, Voinnet O (2009) Roles of plant small RNAs in biotic stress responses. Annu Rev Plant Biol 60:485–510
29. Brigneti G, Voinnet O, Li W-X, Ji L-H, Ding S-W, Baulcombe DC (1998) Viral pathogenicity determinants are suppressors of transgene silencing in Nicotiana benthamiana. EMBO J 17:6739–6746
30. Sansregret R, Dufour V, Langlois M, Daayf F, Dunoyer P, Voinnet O, Bouarab K (2013) Extreme resistance as a host counter-counter defense against viral suppression of RNA silencing. PLoS Pathog 9(6), e1003435
31. Martínez-Priego L, Donaire L, Barajas D, Llave C (2008) Silencing suppressor activity of the Tobacco rattle virus-encoded 16-kDa protein and interference with endogenous small RNA-guided regulatory pathways. Virology 376:346–356
32. Uhrig JF, Canto T, Marshall D, MacFarlane SA (2004) Relocalization of nuclear ALY proteins in the cytoplasm by the Tomato bushy stunt virus P19 pathogenicity protein. Plant Physiol 135:2411–2423

33. Lai H, Chen Q (2012) Bioprocessing of plant-derived virus-like particles of norwalk virus capsid protein under current good manufacture practice regulations. Plant Cell Rep 31:573–584
34. Fullner KJ, Nester EW (1996) Temperature affects the T-DNA transfer machinery of *Agrobacterium tumefaciens*. J Bacteriol 178:1498–1504
35. Chellappan P, Vanitharani R, Ogbe F, Fauquet CM (2005) Effect of temperature on geminivirus-induced RNA silencing in plants. Plant Physiol 138:1828–1841
36. Qu F, Ye X, Hou G, Sato S, Clemente TE, Morris TJ (2005) RDR6 has a broad-spectrum by temperature-dependent antiviral defense role in *Nicotiana benthamiana*. J Virol 79:15209–15217
37. Szyttia G, Silhavy D, Molnár A, Havelda Z, Lovas A, Lakatos L, Bánfaldi Z, Burgyán J (2003) Low temperature inhibits RNA silencing-mediated defence by the control of siRNA generation. EMBO J 22:633–640
38. Velázquez K, Renovell A, Comellas M, Serra P, García ML, Pina JA, Navarro L, Moreno P, Guerri J (2010) Effects of temperature on RNA silencing of a negative-stranded RNA plant virus: *Citrus psorosis virus*. Plant Pathol 59:982–990
39. Adams MJ, Antonew JF (2006) DPVweb: a comprehensive database of plant and fungal virus genes and genomes. Nucleic Acids Res 34(database issue):D382–D385
40. Rizzo TM, Palukaitis P (1990) Construction of full-length RNA clones of cucumber mosaic virus RNAs 1, 2 and 3: generation of infectious RNA transcripts. Mol Gen Genet 222:249–256
41. Carrington JC, Cary SM, Dougherty WG (1988) Mutational analysis of Tobacco etch virus polyprotein processing: cis- and trans- proteolytic activities of polyproteins containing the 49-kDa proteinase. J Virol 62:2313–2320
42. Ahlquist P, French R, Janda M, Laoesch-Fries LS (1984) Multicomponent RNA plant virus infection derived from cloned viral cDNA. Proc Natl Acad Sci U S A 81:7066–7070
43. Donson J, Kearney CM, Hilf ME, Dawson WO (1991) Systemic expression of a bacterial gene by a tobacco mosaic virus-based vector. Proc Natl Acad Sci U S A 88:7204–7208
44. Johnson J, Lin T, Lomonossoff G (1997) Presentation of heterologous peptides on plant viruses: genetics, structure, and function. Annu Rev Phytopathol 35:67–86
45. Igarashi A, Yamagata K, Sugai T, Takahashi Y, Sugawara E, Tamura A, Yaegashi H, Yamagishi N, Takahashi T, Isogai M, Takahashi H, Yoshikawa N (2009) Apple latent spherical virus vectors for reliable and effective virus-induced gene silencing among a broad range of plants including tobacco, tomato, Arabidopsis thaliana, cucurbits, and legumes. Virology 386:407–416
46. Bedoya L, Martínez F, Rubio L, Darós JA (2010) Simultaneous equimolar expression of multiple proteins in plants from a disarmed potyvirus vector. J Biotechnol 150:268–275
47. Chapman S, Kavanagh T, Baulcombe D (1992) Potato virus X as a vector for gene expression in plants. Plant J 2:549–557
48. Baulcombe DC, Chapman S, Santa Cruz S (1995) Jellyfish green fluorescent protein as a reporter for plant virus infections. Plant J 7:1045–1053
49. Santa Cruz S, Chapman S, Roberts AG, Roberts I, Prior DAM, Oparka KJ (1996) Assembly and movement of a plant virus carrying a green fluorescent protein overcoat. Proc Natl Acad Sci U S A 93:6286–6290
50. Sempere RN, Gómez P, Truniger V, Aranda MA (2011) Development of expression vectors based on *Pepino mosaic virus*. Plant Methods 7:6
51. Medina-Escobar N, Haupt S, Thow G, Boevink P, Chapman S, Oparka K (2003) High-throughput viral expression of cDNA–green fluorescent protein fusions reveals novel subcellular addresses and identifies unique proteins that interact with plasmodesmata. Plant Cell 15:1507–1523
52. Dolja VV, McBride HJ, Carrington JC (1992) Tagging of plant potyvirus replication and movement by insertion of β-glucuronidase into the viral polyprotein. Proc Natl Acad Sci U S A 88:10208–10212
53. González-Jara P, Fraile A, Canto T, García-Arenal F (2009) The multiplicity of infection of a plant virus varies during colonization of its eukaryotic host. J Virol 83:7487–7494
54. MacFarlane SA, Popovich AH (2000) Efficient expression of foreign proteins in roots from tobravirus vectors. Virology 267:29–35
55. Ratcliff F, Martín-Hernández AM, Baucombe DC (2001) Tobacco rattle virus as a vector for analysis of gene function by silencing. Plant J 15:237–245
56. Cañizares MC, Nicholson L, Lomonossoff GP (2005) Use of viral vectors for vaccine production in plants. Immunol Cell Biol 83:263–270
57. Stoger E, Sack M, Fischer R, Christou P (2002) Plantibodies: applications, advantages and bottlenecks. Curr Opin Biotechnol 13:161–166

Part VI

Complex Reconstitution

19 Complex Reconstitution from Individual Protein Modules

Jérôme Basquin, Michael Taschner,
and Esben Lorentzen

Abstract

Cellular function relies on protein complexes that work as nano-machines. The structure and function of protein complexes is an outcome of the specific combination of protein subunits, or modules, within the complex. A major focus of molecular biology is thus to understand how protein subunits assemble to form complexes with distinct biological function. To this end, in vitro reconstitution of complexes from individual subunits to study their assembly, structure and activity is of central importance. With purified individual subunits and sub-modules at hand one can systematically dissect the hierarchical assembly of larger complexes using direct protein-protein interaction assays. Furthermore, activity assays can be carried out with individual subunits or smaller sub-complexes and compared to those of the fully assembled complex to precisely map functional sites and provide a molecular basis for in vivo observations. In this chapter we review methods for protein complex assembly from individual subunits and provide examples of advantages and potential pitfalls to this approach.

Keywords

Protein complex reconstitution • Recombinant protein • Size exclusion chromatography

19.1 Introduction

The inner life of a cell is to a large extent the result of the formation and action of a plethora of supramolecular complexes that carry out a wide range of activities and are assembled from individual protein subunits. Protein complexes come in many flavors and varieties. Whereas some are very stable in nature due to high affinity interactions between subunits, other complexes

J. Basquin (✉) • M. Taschner • E. Lorentzen
Department of Structural Cell Biology, Max Planck Institute of Biochemistry,
Am Klopferspitz 18, Martinsried D82152, Germany
e-mail: basquin@biochem.mpg.de

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_19

assemble transiently to fulfill specific spatiotemporal functions. Intraflagellar transport (IFT) "trains", for example, require cycles of dynamic assembly and disassembly to fulfill their function in cilium formation [1, 2]. Similarly, the nuclear pore complex serving as the gate for nucleo-cytoplasmic transport is disassembled into sub-complexes and reassembled during cell division [3]. Yet other assemblies such as RNA degrading exosomes are composed of stable core structures that interact dynamically with regulatory complexes that modulate activity [4]. As a rule of thumb, the reconstitution of stable 'high-affinity' complexes is often relatively straightforward whereas it is much more time consuming to establish conditions that allow for the formation of weakly associated protein assemblies.

Protein domains are classified into ~1000 different folds and ~10,000 different types of protein-protein interactions are estimated to occur in the cell [5]. The evolution of protein complexes to increase the complexity of protein subunit composition has been a key driver of biological function. This is exemplified by the ribosome where the bacterial 70S complex contains 52 protein subunits in addition to ribosomal RNA whereas the eukaryotic 80S counterpart is expanded to 79 protein subunits [6]. Another example of increased complexity is the RNA degrading RNase PH ring that in bacteria contains six identical subunits, in archaea three of each of two different subunits and in eukarya six different but structurally similar subunits [7]. This type of evolution, which is the result of gene duplication followed by mutation, appears to have occurred for many nano-compartments that enclose substrates for degradation or assisted folding such as the proteasome and GroEL-like complexes [8]. The increase in subunit complexity is accompanied by novel functionality such as the ability of the immunoproteasome to produce peptide antigens for surface presentation on antigen-presenting cells of the adaptive immune system [9].

Another important concept when discussing protein complexes is that of modularity [10]. Proteins themselves can be modular when consisting of multiple domains each with specific properties. However, modularity within protein complexes often arises from the close association of protein subunits with different activities. This is exemplified by the ubiquitin ligation system where the E3 ubiquitin ligase simultaneously binds an activated (i.e., ubiquitin-loaded) E2 enzyme as well as a substrate protein, thereby mediating substrate ubiquitylation [11]. A prime example here is the Anaphase Promoting Complex/Cyclosome (APC/C), a 1.5 MDa complex consisting of at least a dozen different subunits, which plays various essential roles in cell cycle progression by ubiquitylating numerous cell cycle regulators and targeting them for proteasomal degradation. The APC/C needs to ubiquitylate specific sets of substrates at well-defined points in the cell cycle, and this temporal specificity is provided by the binding to various co-activators with distinct substrate specificities, the best-studied ones being Cdh1 and Cdc20 [12]. The modularity of proteins has provided nature with efficient means of combining existing modules to achieve new functionality.

Over the last decades, cell biology has transformed from the science of assigning function to individual genes/proteins to a discipline that addresses the function of proteins within the context of larger complexes. In this endeavor, in vitro reconstitution of complexes from individual modules to rigorously test the contribution of each subunit to complex activity is of paramount importance. In this chapter we review and discuss advantages and limitations to protein complex reconstitution from individual subunits.

19.2 Quality Standards for Protein Subunits

19.2.1 Quality Control of Recombinantly Purified Protein Subunits

The reconstitution of protein complexes in vitro relies on the availability of purified protein subunits. The production of recombinant protein using bacterial or eukaryotic expression hosts is covered elsewhere within this book. Before initiating reconstitution experiments it is important to verify that the protein subunits under investigation are soluble and properly folded. Although completely insoluble proteins are typically found in inclusion bodies during over-expression and thus detected at an early stage, apparently soluble protein can form non-functional soluble aggregates of various sizes. Such aggregates can be a result of hydrophobic surfaces that may be exposed when not interacting with other subunits within the context of a protein complex. Soluble aggregation may also result from incorrectly formed disulfide bridges or weak non-covalent interactions. Whereas larger aggregates (0.1–1 mm) can often be detected by eye as a white cloudy suspension, smaller aggregates are best detected using biophysical/biochemical methods such as size exclusion chromatography (SEC). In SEC, aggregates with a size of 0.1–100 μm will elute at or close to the void volume of the column. As SEC often constitutes the final step of protein purification, particular care should be taken in the evaluation of the elution profile to ensure that the sample is not aggregated. In cases of partial aggregation, the soluble fractions can be pooled separately from the aggregation fractions and used in subsequent experiments. In cases where protein subunits prone to aggregation are stored on ice or at −80 °C, it is recommended to repeat SEC to assess if aggregation is induced by storage and to remove any newly formed aggregates. Carrying out experiments with soluble aggregates may prevent complex formation, lead to non-functional complexes and has in some cases been shown to produce artifactual protein interactions and should be avoided [13] (Fig. 19.1). Knowing where individual subunits elute in SEC also allows for the assessment of proper protein complex formation by monitoring an elution shift towards higher molecular weights when individual subunits are mixed to form a complex (Figs. 19.1 and 19.3).

19.2.2 Measures to Prevent Protein Aggregation

In cases where the protein subunit of interest is completely insoluble and forms inclusion bodies, it is often advisable to test alternative expression conditions. Among the most common methods in this regard is to regulate the expression temperature, for example by switching from 37 to 18 °C before induction of protein expression, thus allowing the protein of interest to be produced more slowly. In cases where a eukaryotic protein is insoluble when expressed in *E. coli*, this might suggest a requirement for eukaryotic protein chaperones for proper folding and it may thus be advisable to try eukaryotic expression hosts such as insect or mammalian cells. In cases where the protein of interest forms soluble aggregates, different measures can be taken to alleviate this problem. In cases of non-native disulfide bridges, reducing agents such as β-mercaptoethanol, DTT or TCEP can be added to the buffer or the concentration of such agents increased (it is always instructive to scan the protein sequence for cysteine residues to assess if oxidation may be a problem). Aggregations due to non-covalent hydrophilic interactions may be circumvented by changes in pH or salt concentration of the buffer. Such aggregates may furthermore only manifest themselves at higher protein concentrations and can in some cases be avoided by working below the critical concentration. The formation of soluble aggregates because of exposed hydrophobic surfaces is often the most difficult to deal with. Sometimes the addition of 10–20 % glycerol or possibly detergent to the buffer may be sufficient to overcome the problem. In other cases, the only way to obtain soluble material of such protein subunits may be through the co-expression with the binding partner effectively shielding the hydrophobic surface patch.

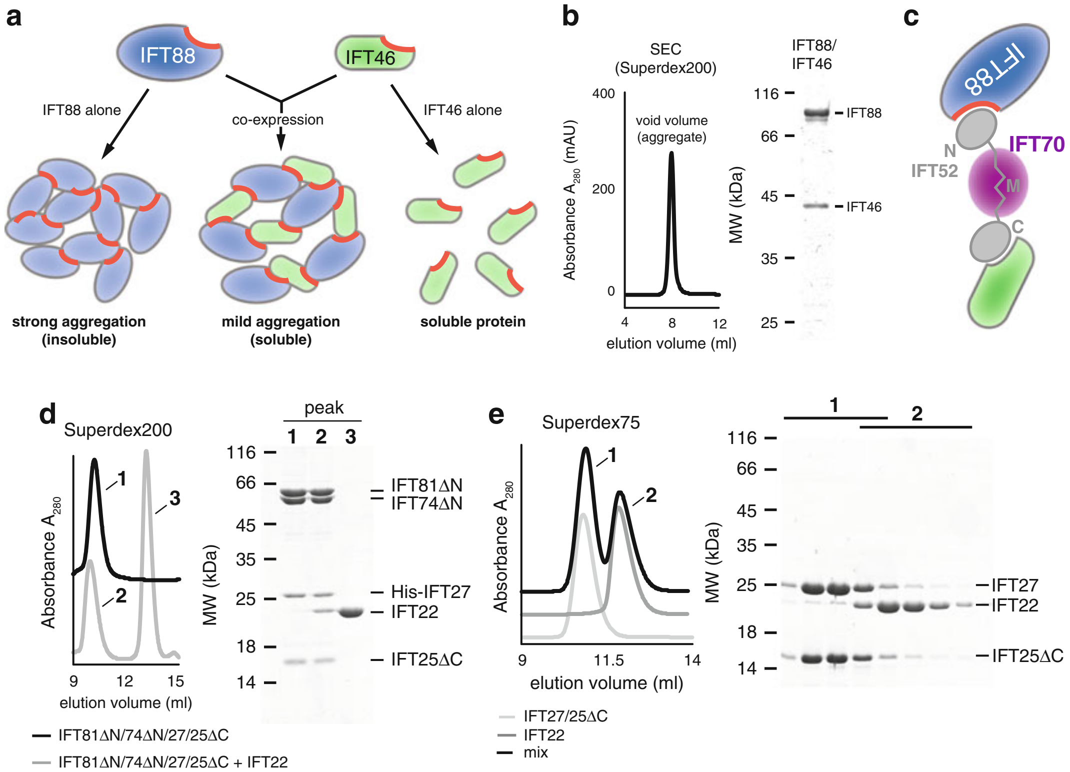


Fig. 19.1 Analysis of complex formation and protein aggregation using Size-exclusion chromatography (*SEC*) (for details see Taschner et al. [13]). (**a**) An example where protein aggregation mistakenly suggests complex formation. IFT88, a member of the Intraflagellar Transport (*IFT*) complex, strongly aggregates and is rendered insoluble when expressed in isolation (*left part*) due to the presence of a hydrophobic surface patch (indicated by a *red line*). IFT46, another member of the complex, can be purified as a soluble protein (*right part*) despite also having a hydrophobic surface. Co-expression of both proteins followed by pull-down experiments suggests complex formation between the two factors, while in fact they form soluble aggregates (*middle part*). (**b**) SEC and SDS-PAGE analysis of the material obtained after IFT88/IFT46 co-expression (*middle part* in **a**) proves that the 'complex' is a soluble aggregate containing both proteins, as it elutes in the void volume of the column (8 ml) suggesting a molecular weight in the MDa range. (**c**) Detailed analysis of the interactions within the IFT complex revealed that the interaction between IFT88 and IFT46 is indirect and mediated by another factor, namely IFT52. IFT88 binds to an N-terminal (N) and IFT46 to a C-terminal (C) domain in IFT52, respectively. These domains are separated by an extended 'middle' domain (M), which is bound by another protein, IFT70. (**d**) Example of SEC and SDS-PAGE analysis showing complex formation between proteins. A tetrameric complex (IFT81/74/27/25, peak 1) clearly elutes at a different volume than another single protein (IFT22, peak 3). After mixing, a stoichiometric amount of IFT22 is pulled into a high-molecular weight (peak 2), which is shifted towards a lower elution volume compared to peak 1. (**e**) Example of SEC and SDS-PAGE analysis showing that two factors do not form a complex. A dimeric complex (IFT27/25, peak 1) elutes at a different volume than another protein (IFT22, peak 2) when mixed. The two peaks overlap perfectly with the elution volumes of the individual factors (©Taschner et al. [13])

19.3 Ligands Required for Protein Folding/Stability

Many proteins require the binding of small molecules or ions for activity and in some cases also proper folding and stability. Examples of this are Zn-finger and EF-hand proteins that rely on the coordination of Zn^{2+} and Ca^{2+} ions, respectively, to adopt their native structure. GTPases and ATPases may require the association with nucleotides, and enzymes often require a co-factor such as NAD, NADP, Acetyl-CoA or FAD to mention a few. In cases where such a dependency is known from previously published studies or is suspected based on sequence homology, it is advisable to include the interactor in the buffer or even into the expression medium. If the protein of interest can be successfully produced and purified without the addition of the small molecule this is sometimes advantageous as it allows for the impact of the ligand to be specifically tested in subsequent in vitro binding or activity assays. Ligands sometimes bind proteins with high affinity (pM–nM dissociation constants) and protein-ligand complexes may form during expression and remain associated during the entire purification. Such cases may result in the identification of novel ligands and sometimes even in the structure solution of the protein subunit bound to a co-purified substrate such as the RNase II enzyme bound to RNA [14]. It is thus advisable to monitor the preparation of the protein subunit of interest for potential association with small molecules using light absorption, mass spectrometry or other biophysical methods.

19.4 Dissecting the Hierarchical Composition of Complexes

A significant advantage of having the individual constituents of a protein complex available as purified components is the possibility to directly map interactions and activities in detail. In this strategy, the different subunits or modules of a complex are produced separately and it is relatively straightforward to perform a multifactorial expression screening where the influence of parameters such as strains, growth conditions, expression helper tags or truncation constructs could be tested. This initial expression screening helps to select the best possible condition to produce the subunits individually. Often the components of a protein complex may be known from other research methods (*e.g.*, co-immunoprecipitation and mass spectrometry) but the direct interactions between subunits are unknown. Once purified components are available, pull-down assays are useful to confirm the existence of protein-protein interactions and to produce a domain-resolution architectural map of the complex of interest. In this assay, a bait protein is tagged and captured on an immobilized affinity ligand specific for the tag; the immobilized bait is then incubated with a protein source that contains putative “prey” proteins. Very often the single subunits are recombinantly produced and harbor a cleavable affinity tag. In this configuration, binary or ternary interactions could be easily tested in pull-down assays where multiple combinations can be tested in parallel to rapidly map interactions. Furthermore, interaction mapping by pull-downs often results in the elucidation of sub-complexes within a larger complex. Such sub-complexes can then be further used in structural studies or activity assays to pinpoint functionality.

Size exclusion chromatography (SEC) has been proven to be a very powerful tool to reconstitute protein complexes. In brief SEC is a separation technique based on the molecular size of the components. The separation of the sample molecules is achieved by the differential exclusion from the pores of the packing material, of the sample molecules as they pass through a bed of porous particles. The principle feature of SEC is its gentle non-adsorptive interaction with the sample. Proteins are prone to interact with surface charged sites of chromatographic stationary phases. These ionic interactions can result in adsorption of the protein, shifts in retention time, peak tailing or peak asymmetry, or to changes in the three dimensional conformation of the protein. A common approach to reduce electrostatic interactions in SEC involves increasing the ionic strength or salt concentration of the mobile phase. This can reduce secondary interactions and improve peak symmetry, retention time, and quantitation. In the frame of protein complex reconstitution, this factor has to be considered carefully; indeed some protein complexes can feature weak interactions and can fall apart at high or moderate salt concentration. In this context it is crucial to screen a range of salt concentration to find a balance between non-absorption to the bed surface and a conservation of the complex integrity. This parameter can be very informative and can give an empirical idea of the stability of the protein complexes. In the context of structural biology studies where sample quality and homogeneity is crucial, SEC is often used as the final step of purification. This last step is essential to remove traces of aggregation in the protein sample.

Reconstitution of protein complexes from individually purified components assisted by SEC is an experiment relatively easy to set up. Each component will be injected separately, their peak retention time will be measured and the peak fractions will be analyzed on SDS-PAGE gels (Figs. 19.1 and 19.3). In the simple case of a heterodimer each monomer will be mixed with a ratio of 1:1.2 M for the smaller component. The mixture will be injected on SEC and analyzed. Upon complex formation the main retention peak will be shifted as compared to the single component and the two proteins should be detectable in the SDS-PAGE gel from the same retention peak. The same strategy can be expanded to more components. In principle there is no upper limit as long as the size of the target complex is compatible with the resolution of the size exclusion column. Advantages to SEC are the facts that stoichiometric complexes, that are often hard to achieve by simply mixing components, can be produced and that the elution volume allows for the molecular weight of the complex to be estimated, although caution is warranted as the shape of the complex will influence the retention time on the column.

In a situation of multi-modular protein complexes, SEC can be used to decipher the network of interactions between the different components of the complex and map the different modules. This strategy can be further refined by combining SEC and limited proteolysis. Limited proteolysis is a method based on the proteolytic susceptibility of a specific, exposed flexible chain in a folded protein. Complexes can be subjected to limited proteolysis using several proteases. The resulting proteolytic products can be analyzed by SDS-PAGE, and the integrity of truncated sub-complexes confirmed by analytical SEC. The exact proteolytic fragments of each subunit within a stable sub-complex can be determined by mass spectrometry. This approach is of particular interest when aiming to understand the network of interactions at domain-resolution or to reconstitute smaller sub-complexes for activity assays or structure determination. This approach has been used to characterize the IFT70/52/46 complex [15] and is illustrated in Fig. 19.2.

19.5 Coupling Co-expression and In Vitro Reconstitution Strategies

Once the interaction network of a complex is established, it is possible to design protein expression strategies for larger scale production. Typically structural biology studies require several mg of pure complexes. For large complexes it is often advantageous to co-express and purify modules or sub-complexes separately. This approach is often necessary to overcome the limitation of expressing very large complexes in heterologous systems. It is also possible to use different expression systems for the separate modules depending on their expression success in a given expression system. For example a complex can be reconstituted from protein modules expressed in *E. coli* and insect cells, respectively. The modules can then later be mixed and the final complex reconstituted and purified by SEC. This approach has been used to reconstitute the nuclease module of the yeast CCR4-NOT complex [16] and is illustrated in Fig. 19.3. This modular approach is very powerful and flexible as it combines the advantages of using different expression systems and can also accommodate time constrains from the multi-step purification process.

19.6 Dissecting the Activities of Subunits and Complexes

One major advantage to the method of complex reconstitution from individual subunits is that the activities of the complex can be rigorously mapped given a reliable in vitro activity assay. With individually purified components in hand, one can undertake a systematic approach where the activities of individual components as well as all possible combinations of sub-complexes are

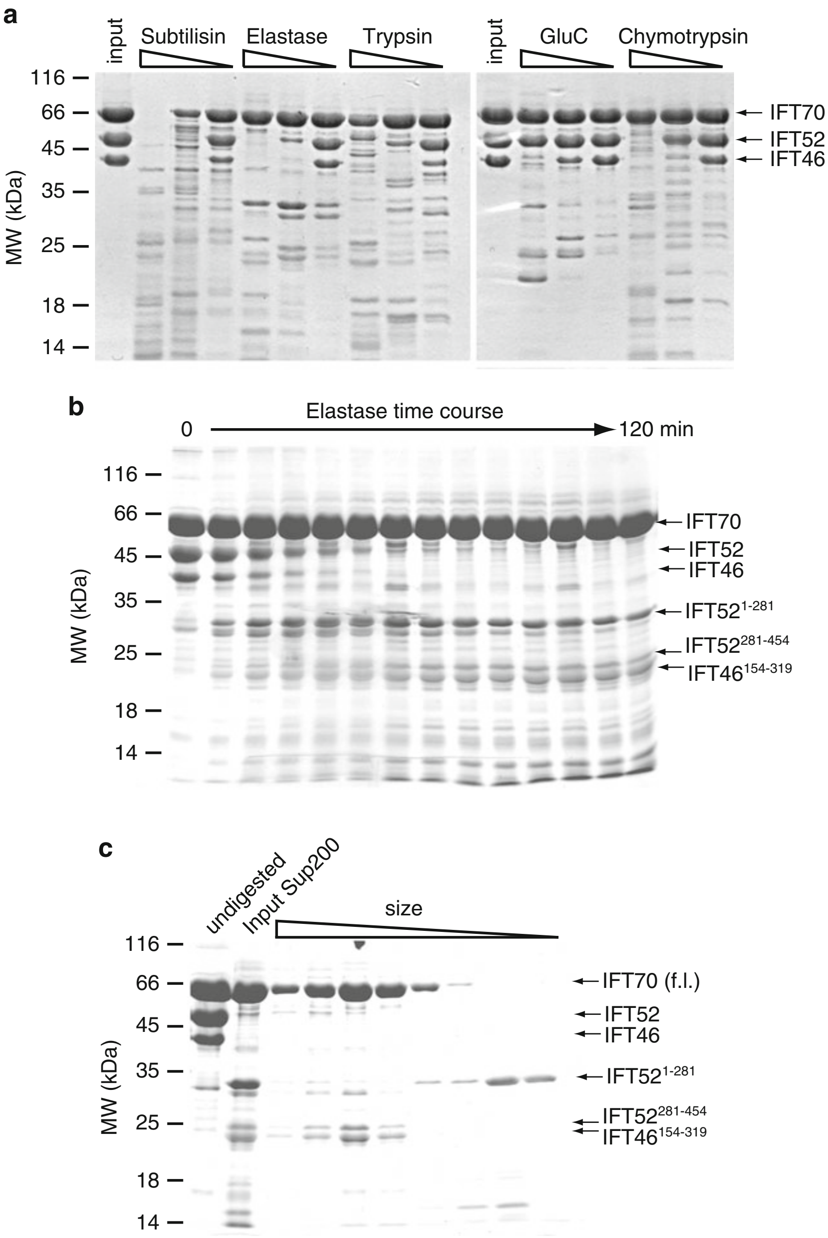


Fig. 19.2 Limited proteolysis carried out on the IFT70/52/46 complex. (**a**) In the initial small-scale test a small volume of diluted protein complex (typically 10 μl of around 1 mg/ml) is incubated with varying amounts of several proteases for a defined time at a certain temperature. SDS-PAGE analysis is then used to check which protease cuts the starting material into defined smaller bands. In this example IFT70 is relatively stable whereas both IFT52 and IFT46 are cleaved into discrete bands by Elastase. (**b**) In the next step a concentrated protein complex (*e.g.*, 20 mg/ml) is incubated with a certain amount of the protease identified in the initial small-scale screen. Samples are taken out in short intervals (*e.g.*, every 10 or 20 min) and analyzed by SDS-PAGE to find the time point at which the desired stable fragments have been formed. (**c**) In the final step the reaction from step **b** is scaled up to produce enough material suitable for SEC analysis. After incubation of the complex and the protease for the time determined in step **b** the material is directly loaded on a pre-equilibrated size-exclusion column, and co-migrating fragments are identified by mass spectrometry (This research was originally published in the Journal of Biological Chemistry (Taschner et al. [17]) © the American Society for Biochemistry and Molecular Biology)

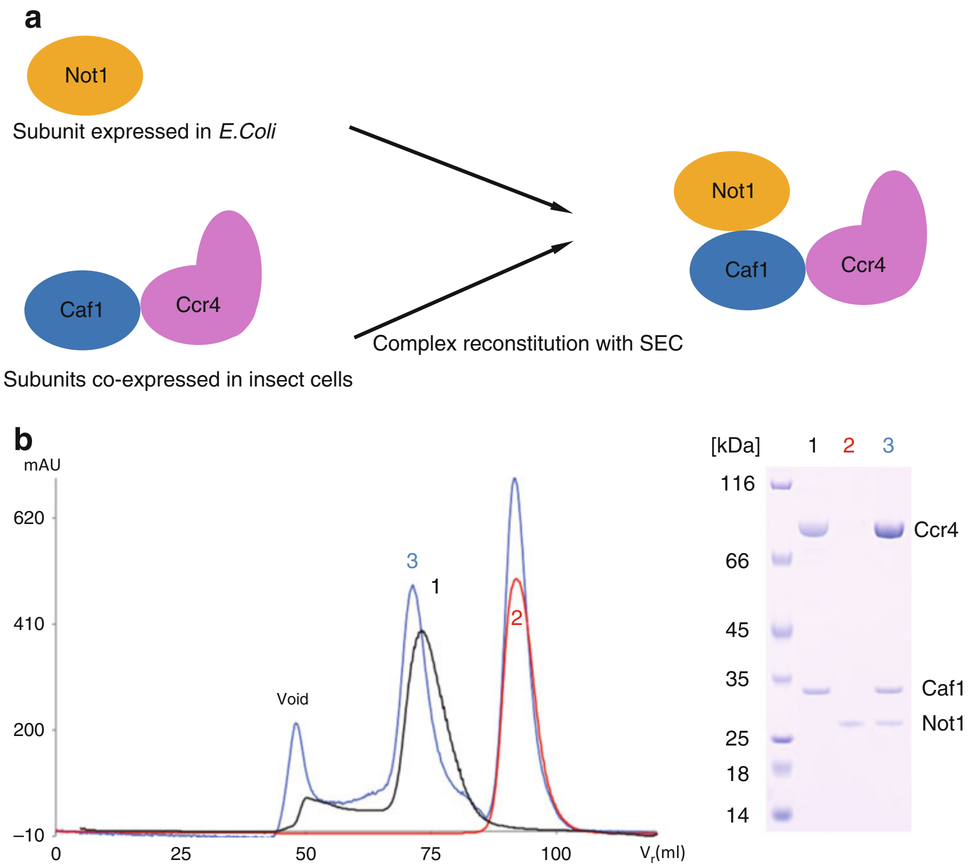


Fig. 19.3 Reconstitution of the nuclease module of the Ccr4-Not complex. (**a**) Scheme of the reconstitution strategy. (**b**) Ccr4 was co-expressed with Caf1, purified and analyzed by size exclusion chromatography in the absence (peak 1) or presence (peak 3) of the CNOT1 MIF4G domain (Superdex S200 Hiload 16/60 column, GE Healthcare). On the left are the overlays of the chromatograms (mAU and Vr denote relative absorbance and retention volume of the proteins, respectively). On the right is the Coomassie stained SDS-PAGE gel with the samples from the corresponding peak fractions

tested. This notion is illustrated in Fig. 19.4 by the archaeal exosome complex involved in RNA degradation and processing [4]. The archaeal exosome core represents a simple system composed of only the two subunits Rrp41 and Rrp42 that associate to form a hexameric ring [7]. To map the activity of the archaeal exosome, RNA degradation assays were undertaken using either purified Rrp41, Rrp42 or the Rrp41–42 complex (Fig. 19.4). Interestingly, although the archaeal exosome is known to be an active RNase (Fig. 19.4b, lane 1), neither the Rrp41 nor the Rrp42 subunit alone displayed any catalytic activity when incubated with an RNA substrate (Fig. 19.4b, lanes 5–6), which demonstrate that the pre-formation of a complex is a requirement for functionality. The molecular rationale for this observation was obtained from the crystal structure of Rrp41–42 in complex with RNA [7] demonstrating that whereas the Rrp41 subunit harbors the catalytic site, Rrp42 harbors critical RNA substrate binding residues required for the recognition of the substrate (Fig. 19.4a). Another advantage to this approach is that point mutations can be introduced into the recombinantly expressed subunits to pinpoint catalytic and substrate binding residues (Fig. 19.4b, lanes 2 and 4). Similarly, the importance of individual domains can be tested by expressing and purifying truncated versions of one or more subunits within a

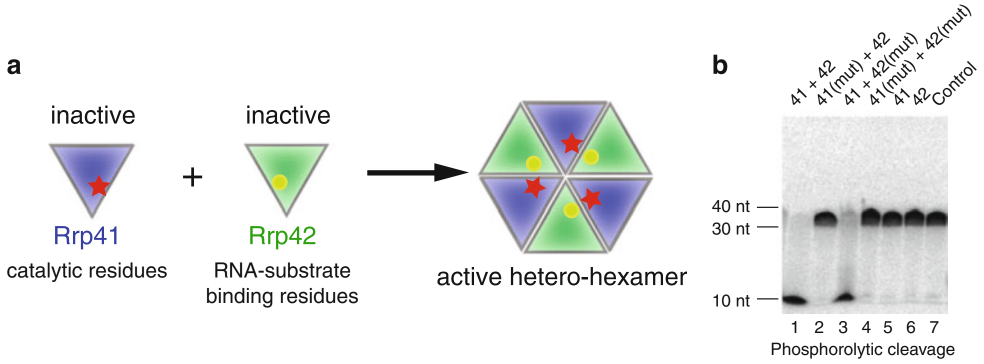


Fig. 19.4 Activity assay of reconstituted complex. Activity assay of reconstituted complex: (**a**) Schematics of the Rrp41 subunit containing active site residues and the Rrp42 subunit containing substrate-binding residues. The association of Rrp41 and Rrp42 results in a hexameric ring structure bringing substrate-binding and catalytic residues into close proximity. (**b**) RNA degradation assay showing a gel of an RNA-substrate (*upper band*) and the resulting degradation fragment (*lower band*). Whereas neither Rrp41 nor Rrp42 have any RNA degradation activity (lanes 5–6), the Rrp41-Rrp42 complex is an active RNase (lane 1) (**b** was initially published in *Nature Structural and Molecular Biology* (Lorentzen [18]), ©Nature Publishing Group)

complex. Furthermore, once an activity assay with reconstituted components is established, it is relatively straightforward to assay for the impact of newly identified subunits or modulators of a given complex. Complex formation from individual subunits clearly provides an advantage to an entirely co-expressed complex in the sense that the dissection of functional sites and subunit-interdependent functionality can readily be assayed and should be utilized whenever possible.

19.7 Conclusions

Reconstitution of protein complexes from individual components has a number of advantages. It allows (1) for the hierarchical assembly of sub-complexes in addition to the full complex, (2) for the detailed mapping of interactions between subunits, (3) for the rigorous testing of activities of individual subunits and sub-complexes to pinpoint active sites, and (4) for the structural analysis of subunits and sub-complexes in cases where the fully assembled complex does not crystallize. However, there are also a number of disadvantages and pitfalls. The task of producing individual subunits for very large complex (for example the ribosome) may simply be too laborious and in such cases native purification of the entire complex might be a better choice. The largest obstacle to complex reconstitution will generally be difficulties in obtaining soluble recombinant material for all implicated subunits and often a considerable amount of time may have to be invested in optimizing protein production. Furthermore, special care has to be taken in evaluating that the individual protein subunits produced are properly folded and functional. However, for most protein complexes of say 2–15 subunits in size, the flexibility and possibility of dissecting interactions and activities in a rigorous manner may make the method of complex assembly from individual subunits the method of choice.

References

1. Bhogaraju S, Engel BD, Lorentzen E (2013) Intraflagellar transport complex structure and cargo interactions. Cilia 2(1):10
2. Taschner M, Bhogaraju S, Lorentzen E (2012) Architecture and function of IFT complex proteins in ciliogenesis. Differentiation 83(2):S12–S22
3. Rabut G, Lénárt P, Ellenberg J (2004) Dynamics of nuclear pore complex organization through the cell cycle. Curr Opin Cell Biol 16(3):314–321

4. Lorentzen E, Basquin J, Conti E (2008) Structural organization of the RNA-degrading exosome. Curr Opin Struct Biol 18(6):709–713
5. Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. Nat Biotechnol 22(10):1317–1321
6. Brodersen DE, Nissen P (2005) The social life of ribosomal proteins. FEBS J 272(9):2098–2108
7. Lorentzen E, Conti E (2005) Structural basis of 3′ end RNA recognition and exoribonucleolytic cleavage by an exosome RNase PH core. Mol Cell 20(3):473–481
8. Lorentzen E, Conti E (2006) The exosome and the proteasome: nano-compartments for degradation. Cell 125(4):651–654
9. Basler M, Kirk CJ, Groettrup M (2013) The immunoproteasome in antigen processing and other immunological functions. Curr Opin Immunol 25(1):74–80
10. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402(supp):C47–C52
11. Hershko A, Ciechanover A (1998) The ubiquitin system. Annu Rev Biochem 67(1):425–479
12. Peters J-M (2006) The anaphase promoting complex/cyclosome: a machine designed to destroy. Nat Rev Mol Cell Biol 7(9):644–656
13. Taschner M, Kotsis F, Braeuer P, Kuehn EW, Lorentzen E (2014) Crystal structures of IFT70/52 and IFT52/46 provide insight into intraflagellar transport B core complex assembly. J Cell Biol 207(2):269–282
14. Frazão C, McVey CE, Amblar M, Barbas A, Vonrhein C, Arraiano CM, Carrondo MA (2006) Unravelling the dynamics of RNA degradation by ribonuclease II and its RNA-bound complex. Nature 443(7107):110–114
15. Taschner M, Bhogaraju S, Vetter M, Morawetz M, Lorentzen E (2011) Biochemical mapping of interactions within the Intraflagellar Transport (IFT) B core complex. J Biol Chem 286(30):26344–26352
16. Basquin J, Roudko VV, Rode M, Basquin C, Séraphin B, Conti E (2012) Architecture of the nuclease module of the yeast Ccr4-Not complex: the Not1-Caf1-Ccr4 interaction. Mol Cell 48(2):207–218
17. Taschner M, Bhogaraju S, Vetter M, Morawetz M, Lorentzen E (2011) Biochemical mapping of interactions within the Intraflagellar Transport (IFT) B core complex: IFT52 binds directly to four other IFT-B subunits. J Biol Chem 286:26344–26352
18. Lorentzen E et al (2005) The archaeal exosome core is a hexameric ring structure with three catalytic subunits. Nat Struct Mol Biol 12:575–581

20 Structural Reconstruction of Protein-Protein Complexes Involved in Intracellular Signaling

Klára Kirsch, Péter Sok, and Attila Reményi

Abstract

Signaling complexes within the cell convert extracellular cues into physiological outcomes. Their assembly involves signaling enzymes, allosteric regulators and scaffold proteins that often contain long stretches of disordered protein regions, display multi-domain architectures, and binding affinity between individual components is low. These features are indispensable for their central roles as dynamic information processing hubs, on the other hand they also make reconstruction of structurally homogeneous complex samples highly challenging. In this present chapter we discuss protein machinery which influences extracellular signal reception, intracellular pathway activity, and cytoskeletal or transcriptional activity.

Keywords

Cellular signaling • X-ray crystallography • Cryo-EM • Protein-protein interactions • Signaling pathway

20.1 Introduction

Signal transduction refers to all molecular events between the reception of extracellular signals and the mounting of biologically appropriate responses inside the cell (*e.g.*, gene expression by the general transcriptional machinery or movements involving cytoskeletal proteins). As cells receive myriad of signals and responses are functionally diverse, a great proportion of intracellular proteins participate in the hierarchical assembly of signaling complexes. Protein-protein interaction specificity of components within these complexes determines how signaling pathways are wired. We show that detailed mechanistic understanding on how signaling complexes transmit intracellular information requires their structural reconstruction. However, this is difficult, because signaling proteins often form short-lived transient complexes, are prone to allosteric regulatory mechanisms, and modulated by post-translational modifications (*e.g.*, phosphorylation

K. Kirsch • P. Sok • A. Reményi (✉)
MTA "Lendület" Protein Interaction Group, Institute of Enzymology, Research Center for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
e-mail: remenyi.attila@ttk.mta.hu

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*, Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_20

or nondegradative ubiquitinilation). In addition, most proteins contain long disordered protein regions, display multi-domain architecture, and binding affinities between structured and linear motif containing disordered regions are weak (micromolar).

In the next pages we will review how the above mentioned technical challenges were solved for reconstructing GPCR-G protein complexes, focal adhesion sites, the Ste5 MAP kinase cascade, the ARP2/3, and the Mediator complex. The structural reconstruction of these complexes has given insight into the reception of chemical ligands, adhesion to the extracellular matrix, intracellular signaling cascade insulation, actin branching dynamics and transcriptional activation, respectively. These topics give a cross-section of now structurally explored molecular events from the cellular signaling field. On the other hand, the examples below maybe viewed as paradigmatic cases on how to devise strategies to limit conformational flexibility of reconstituted multi-protein complexes, or alternatively to divide them up into functionally relevant and structurally compact units.

20.2 Sensing the Chemical Environment: GPCRs and Heterotrimeric G Proteins

G protein-coupled receptors (GPCRs) play a central role in detecting extracellular signals. They bind ligands outside of the cell, go through binding triggered conformation changes and turn these into downstream intracellular signals with the help of heterotrimeric G-proteins located at the cytoplasmic side of the cell membrane. GPCRs are an important group of signaling receptors as the largest part of current drugs deliver their effects through them. Learning the molecular mechanism of GPCR activation is the key to create successful therapeutics. However, acquiring insight into the conformational changes of GPCRs upon ligand binding has turned out to be a difficult task [1]. The first solved GPCR structure was the light sensitive but relatively stable rhodopsin [2]. Most GPCR proteins are hard to express in the necessary amounts and are unstable when using common detergent solubilization methods. Instead of detergents, most of the GPCR crystals were grown in lipidic cubic phase where proteins are stabilized by the membrane bilayer [3].

The highly dynamic nature of GPCRs has hampered their structural investigation for a long time. Structure solution of the beta-2-adrenergic receptor (β2AR) was made possible by using monoclonal antibodies to stabilize conformation of a flexible region—the third intracellular loop connecting two transmembrane regions, helices 5 and 6 [4]. As an alternative to this, insertion of the stable T4 lysozyme (T4L) protein at this region was also successfully used to stabilize the GPCR structure [5]. In addition, using inverse agonists such as carazolol was helpful in locking the GPCR into its inactive conformation, causing less conformational heterogeneity at other flexible protein regions [6].

Obtaining an active, agonist-bound GPCR structure has also proven difficult due to the inherent instability of this state in the absence of a G protein. This was circumvented by using nanobodies, single domain antibodies that exhibited G protein-like behaviour [7, 8].

Binding of agonists to the extracellular region of the GPCR induces a conformational change in the receptor. The activated GPCR receptor allosterically activates the heterotrimeric G protein. The activated G proteins alpha subunit (Gα) exchanges GDP for GTP, which results in the dissociation of the Gα from the Gβ-Gγ subunits. Activated Gs protein binds and then turns on adenyl cyclase (Fig. 20.1). In 2011 the β2AR-Gs protein complex was finally solved [8]. This was a great contribution to fully understanding the molecular mechanism behind GPCR signaling as well as to know how most drugs exercise their effect. Similarly to efforts on the monomeric GPCR, fighting against and prevailing over conformational heterogeneity of receptor samples was the key for success and several former methodological improvements on how to handle GPCR samples had to be combined. The fusion of the T4L protein as well as the use of a nanobody

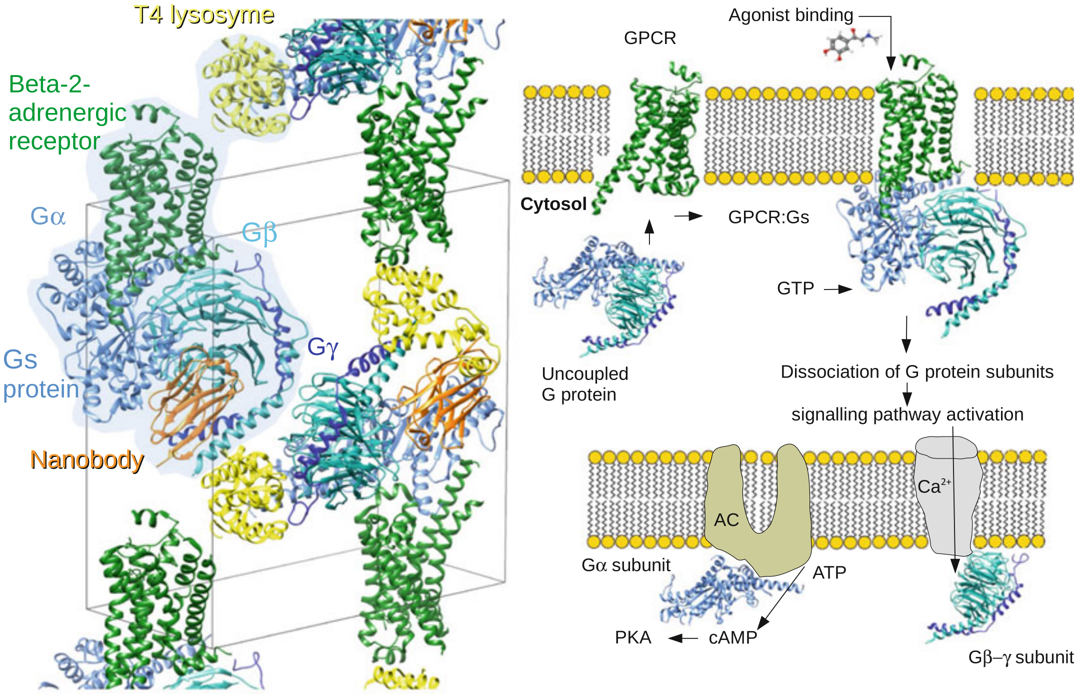


Fig. 20.1 **GPCR structure and signaling**. The panel on the *left* displays the unit cell of the crystal structure of the beta-2-adrenergic receptor and the Gs protein complex. The T4L fusion contributes to the crystal lattice contacts and the nanobody stabilizes a signaling-competent conformation of Gs. One T4L-β2AR-Gs-Nb35 nanobody complex is highlighted in *blue* background. Panels on the *right* show the signaling events after GPCR ligand binding. The activated Gs protein dissociates, its alpha subunit activates adenyl cyclase (AC) and the produced cAMP activates Protein kinase A [49, 50]. The G protein beta and gamma subunit complex also have regulatory functions. It is known to have regulatory effect on calcium ion channels for example [51]

(Nb35) was necessary to stabilize the complex and provide optimal crystal lattice contacts. Finally the T4L-β2AR-Gs-Nb35 protein complex was successfully crystallized in lipidic cubic phase (Fig. 20.1).

Understanding how ligands induce activating conformational changes in GPCRs required some truly creative and novel methods to be applied for their crystallization. Many years of method developments were required to learn how it is pragmatically possible to decrease the inherent flexibility of these dynamic molecular switches.

20.3 Sensing the Matrix: Focal Adhesions

The focal adhesion of cells to the extracellular matrix (ECM) or to neighboring cells is an interesting example for showing how cells could gain information about their physical environment. Focal adhesions are macromolecular assemblies connecting cells to physical surfaces. In adhesion signaling the recruitment of many adaptor proteins to the plasma membrane mediate the outcome of the response. Integrins have a major role in forming focal adhesions and in transducing biochemical signals. Major components of “integrin adhesomes” are paxillin, talin, and vinculin. Overall, they may be composed of more than 150 components and closer examination of this complex network revealed the existence of functional subnets. Key network motifs were dominated by three-component complexes in which a scaffolding molecule recruits both a signaling molecule and its downstream target [9]. Integrin signaling plays a role in cell migration, immune and inflammatory responses, and also in actin polymerization involving the ARP2/3 complex (see later).

The characterization of protein-protein interactions in adhesion contacts are mostly based on Fluorescence Resonance Energy Transfer (FRET), fluorescence co-localization, acceptor photobleaching FRET (apFRET), Fluorescence Recovery After Photobleaching (FRAP) based assays and immunofluorescence imaging [10–12]. These studies revealed interacting proteins at focal adhesion sites. Since many structures of key protein-protein complexes are not known, most molecular mechanisms still remain undiscovered.

Cryo-electron tomography recently gave fundamental insight into the core of focal adhesion sites within cells [13]. Under cryogenic conditions focal adhesions were identified in cells by fluorescent microscopy based on YFP labeled paxillin and by immunolabelled vinculin. The identified components were indexed for cryo-electron tomography. As a result it was possible to identify adhesion related intact integrin-paxilin-vinculin-actin complexes, and their structure could be revealed at ~4–6 nm resolution. As complexes were analyzed in cells, imaging gave information about localization of adhesome particles within the cell. This analysis revealed that the membrane–cytoskeleton interaction at focal adhesions is indeed mediated through particles that are directly attached to actin fibers (Fig. 20.2).

Integrins are heterodimeric receptors of alpha and beta subunits and they are linked to the intracellular cytoskeleton through their short cytoplasmic tails [14, 15]. These cytoplasmic tails are flexible and serve as a hub for adaptor proteins that recruit other interaction partners [16]. Paxillin is one of the well-characterized adaptor protein for integrins which integrates signaling and structural proteins into adhesion sites. It functions as a platform to coordinate multiple signaling pathways and to control the reorganization of the cytoskeleton. One of its major partner is focal adhesion kinase (FAK) which is a central signaling protein recruited to adhesomes. FAK is a multi-domain tyrosine kinase [17]. NMR studies on the interacting domains of FAK (FAT domain) and paxillin (LD motifs) revealed the highly dynamic nature of this important regulatory interaction [18]. Focal adhesions are abundant in regulatory proteins such as protein kinases, phosphatases, GTPases, GAPs, and GEFs. Because these are not only affected by upstream signaling events coming from the receptor but in turn they also modify the receptor, integrin signaling is a two-way signaling process where besides mediating signals from outside to inside, cells could alter their integrin binding affinity to its ligands for inside-out signaling [19] (Fig. 20.2).

Focal adhesions are complex and dynamic structures comprised of high number of protein components. Once protein binding profiles are mapped out, structural investigation of important binary or ternary sub-complexes is possible, however, understanding how they connect integrin receptors to the cytoskeleton will naturally require investigation of at least the core complex in the cell. Cryo-electron tomography on specifically labeled multi-protein containing cellular structures gives unique structural information, albeit at low resolution, which is not possible through reconstituting complexes from purified components *in vitro*.

20.4 Organizing Protein Kinases into Functional Modules

Intracellular signaling pathways often use cascades of protein kinases to mediate signals from the cell membrane. Interestingly, signaling cascades often use shared enzymatic components. At the mechanistic level the question then arises as to how functionally distinct pathway activities are insulated. The solution may be the use of multi-domain scaffolds consisting of dedicated binding proteins capable of assembling different sets of protein kinases. Scaffold proteins potentially allow the combinatorial use of a limited set of signaling enzymes to control a great number of signaling activities [20]. Scaffolds, however, do not merely facilitate signaling between recruited enzymatic components by passive tethering but they also allosterically modulate their bound partners. Recent studies on scaffolds of mitogen activated protein kinase (MAPK) pathways

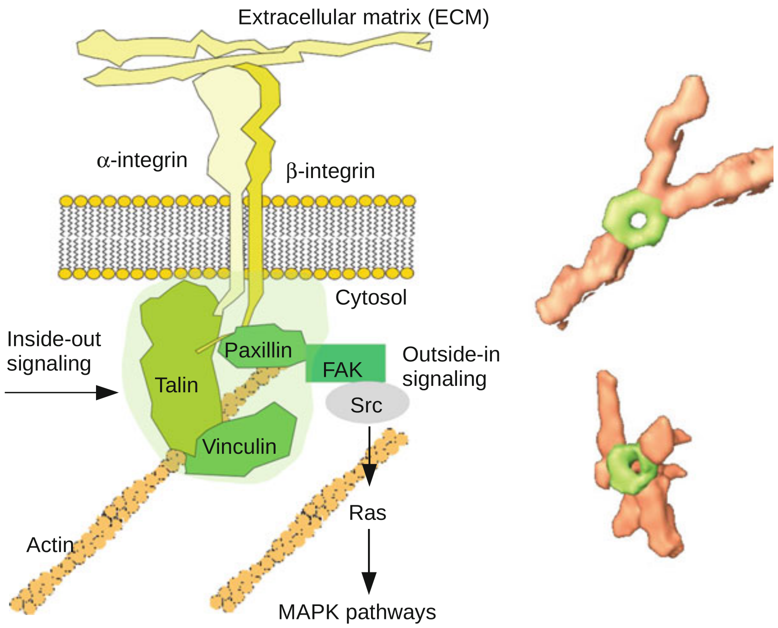


Fig. 20.2 **Focal adhesion sites**. Schematic representation of adhesome particles. The complex shown in green contains the paxilin-vinculin-talin adapter complex that couples integrin receptors to actin and to the focal adhesion kinase (FAK). The recruitment of kinases (*e.g.*, FAK and Src) ensures the functional linkage to downstream signaling pathways (*e.g.,* Ras/MAPK). Besides this outside-in signaling, integrin can be regulated through talin by inside-out signaling. Panels on the *right* show two different adhesome relevant particles comprised of paxilin, talin and vinculin (in *green*) connected to the cytoskeleton (actin in salmon) (Cryo-electron tomography images were taken from Ref. [13])

demonstrated this elegantly by reconstituting scaffolded MAPK modules out of components in well-defined conformational states [21].

One of the best characterized signaling pathway is the baker's yeast α-pheromone response (mating pathway) [22]. This is a classical GPCR triggered pathway that is dependent on an evolutionarily conserved, three-tiered kinase cascade (Fig. 20.3). The three kinases (Ste11, Ste7 and Fus3) sequentially activate each other and can simultaneously bind to the Ste5 scaffold protein. Upon activation of the GPCR the dissociated βγ-subunit of the G-protein recruits Ste5 to the cell membrane, which brings about the activation of the first protein kinase, Ste11, by a membrane located kinase, Ste20. In turn, Ste7 gets activated which will then activate the Fus3 mitogen-activated protein kinase (MAPK). Activated Fus3 enters the nucleus and phosphorylate transcription factors that execute the mating response (where a-type haploid cells fuse with α-type haploid cells to form diploids.) Interestingly, other physiologically non-related pathways also use Ste7 as a common signaling mediator. For example the filamentous growth pathway depends on the Ste7 mediated activation of the Kss1 MAPK. How can Fus3 be selectively activated by Ste7 molecules that obtained upstream signals from the mating but not from the filamentous growth pathway? The answer lies in the Ste5 dependent allosteric activation mechanism of Fus3 by Ste7. In contrast to Kss1, Fus3 can only be activated by Ste7 if it is co-bound with its activator kinase on the Ste5 scaffold [23]. In addition, Ste5 itself is also allosterically regulated. An internal interaction between two of its domains hinders its allosteric role on Ste7-Fus3 signaling, while this is relieved upon its membrane recruitment following GPCR activation [24]. These mechanisms ensure that Ste7 can be used in two unrelated pathways in a physiologically relevant fashion.

Scaffold proteins are abundantly used in MAPK signaling pathways [25]. Similar recon-

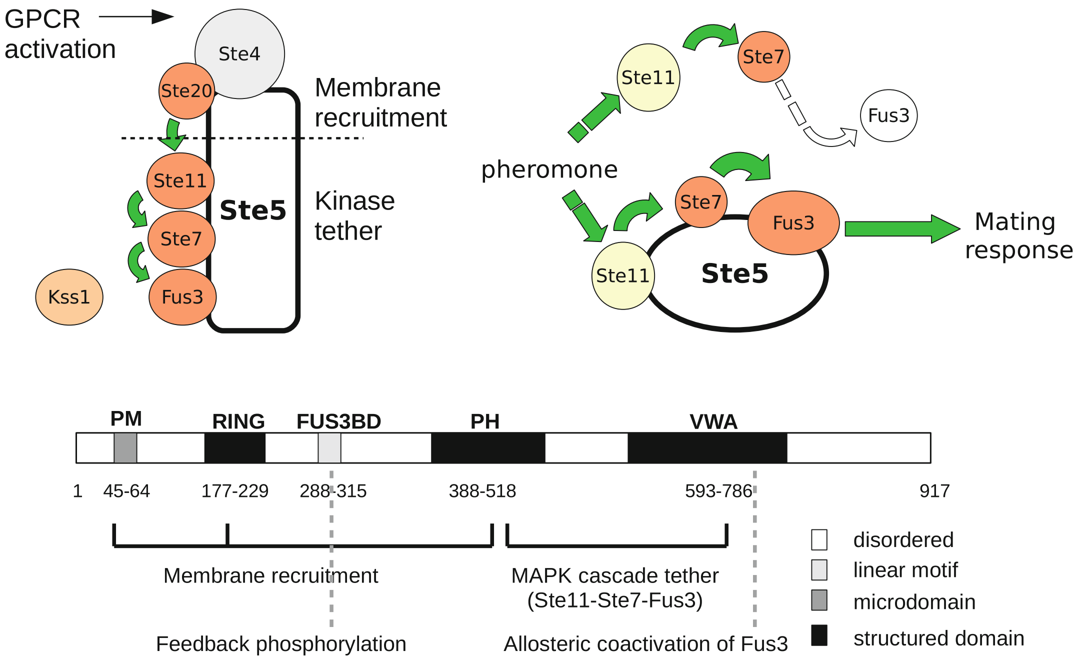


Fig. 20.3 **Modular interactions of the Ste5 scaffold**. Ste5 contains close to 1,000 amino acids. Long stretches of disordered protein regions are interspersed with differently structured regions. PM is an amphipatic alpha helix that binds membranes, the RING domain binds to Ste4 which is the β-subunit of the heterotrimeric G-protein, the Fus3 binding domain (FUS3BD) is a linear motif that adopts a defined conformation only when it is bound to the Fus3 kinase, the pleckstrin homology (PH) domain binds to membrane phosphoinositides, and von Willebrand type A (VWA) domain binds Ste7. Some of these regions play a role in the core steps of signal propagation through the scaffolded complex (*e.g.*, membrane recruitment, tethering MAPK cascade components and allosteric coactivation of the MAPK), while others are involved in higher-order regulatory mechanisms (*e.g.*, negative regulation of membrane recruitment by other kinases, PM; or feed-back phosphorylation by the activated MAPK, FUS3BD)

stitution studies as described above with MAPK module components of the epidermal growth factor sensing pathway also highlighted the importance of allosteric regulation and the existence of multiple, dynamic conformational states. This pathway culminates in the activation of the mammalian MAPK homolog of Fus3, ERK2, and it contains the three-tiered Raf-MEK-ERK module where the KSR scaffold plays somewhat analogous functions to that of Ste5. Here structural studies on sub-complexes of this module showed that KSR-Raf heterodimerization results in an increase of Raf-induced MEK phosphorylation via the KSR-mediated relay of a signal from Raf to release the activation segment of MEK for phosphorylation [26].

Scaffold proteins are normally multi-domain proteins comprised of folded domains and linear motifs with long stretches of disordered protein regions linking these together. Their bound enzymatic components and even scaffolds themselves are subject to function modifying modifications as well as to mutual allosteric regulation. These make the reconstitution of complete scaffolded modules in well-defined functional states technically impossible. The main problem is that these complexes even if reconstituted from homogenous protein sample components, they are too flexible, and thus too heterogeneous for any single particle cryo-EM or crystallography based approaches, and far too big for NMR. Thus researchers have used the "divide and conquer" strategy and focused on characterizing the nature of binary interactions between scaffold-kinase and kinase-kinase pairs. The mechanistic understanding on how the functionally meaningful scaffolded module works comes from by piecing together data obtained on sub-complexes.

20.5 Controlling Cytoskeletal Structure and Dynamics

The dynamic polymerization, depolymerization and branching of actin filaments are controlled by more than a hundred actin-binding proteins [27]. How upstream signals influence this complex network? One of the most studied regulator complex is the actin-related protein-2/3 (ARP2/3) complex, which is responsible for the formation of branched actin filaments. Structural reconstitution experiments seek to reveal the regulatory mechanism of this complex in order to better understand its role in various processes from cell migration, endocytosis, vesicle trafficking, cytokinesis to tumor-cell invasion and metastasis [28]. ARP2/3 is a stable complex of seven conserved subunits (Fig. 20.4). ARPC2 and ARPC4 form the structural core of the complex, ARP2 and ARP3 are involved in the nucleation process, and ARPC1, ARPC3, and ARPC5 contribute to the activation of the complex by N-WASP (neuronal Wiskott-Aldrich Syndrome Protein). Upstream activators responsible for actin regulation (*e.g.*, Cdc42-GTP and PIP2) can bind N-WASP, which disrupts its auto-inhibitory intramolecular interaction. The unmasked VCA domain can bind ARP2 and ARP3 subunits, and branching will be started by binding to the mother actin filament. The pseudo actin dimer composed of ARP2 and ARP3 act as a template for the building of the new filament joining to its mother with 70° Y angle [29, 30].

The reconstitution of human recombinant ARP2/3 complex provided insights into the role of the individual subunits on the stability of the complex as well as on the nucleation of branched filaments [31]. During the reconstitution of any complex it is necessary to fix the conformational states of the monomers to gain a homogeneous sample. In the case of transient interactions it is particularly challenging to determine the conditions for capturing the complex in its active state. For the ARP2/3 complex, several studies tried to resolve the inactive and active states. Beyond crystallization—which has the limitation of freezing the complex into only one state—cryo-EM has been applied to follow transitions between different molecular states [30]. *In vitro* reconstitution of the active ARP2/3 complex is a multi-step process. The active conformation is the result of a conformational change that brings ARP2 and ARP3 subunits together mimicking two sequential subunits in an actin filament (Fig. 20.4). This process requires many components: ATP, Mg, N-WASP, mother actin filament and G-actin monomers.

The first solved crystal structure was the bovine ARP2/3 complex in its inactive state [32]. The architecture confirmed the structural similarities of ARP2 and ARP3 with actin as well as the central role of the core proteins ARPC2 and ARPC4. Homology modeling showed that the important contact points and residues are evolutionary all conserved [33]. *In vitro* FRET studies proved that binding both the nucleotide and NPF is essential for the formation of active ARP2/3 complex [34]. YFP or GFP labeling of individual subunits of the complex enabled their docking into electron micrographs obtained on reconstituted branched actin [35].

Three conformational classes of particles were discovered on the EM grids of wild type yeast and bovine ARP2/3 samples [36]. The open, intermediate and closed states imply great structural flexibility. Further examination revealed that the cryo-EM maps changed when the complex bound to regulator molecules: the inhibitor coronin bound to the ARPC2 subunit and stabilized the open complex, while the activator N-WASP locked it into the closed (active) conformation.

Complex regulatory machines may exist in multiple conformational states and structural reconstitutions first should target only the core part responsible for setting up the basic architecture of the complex. Later, including components outside the core is not only necessary to mechanistically understand activation but also to stabilize conformations that represent important functional states.

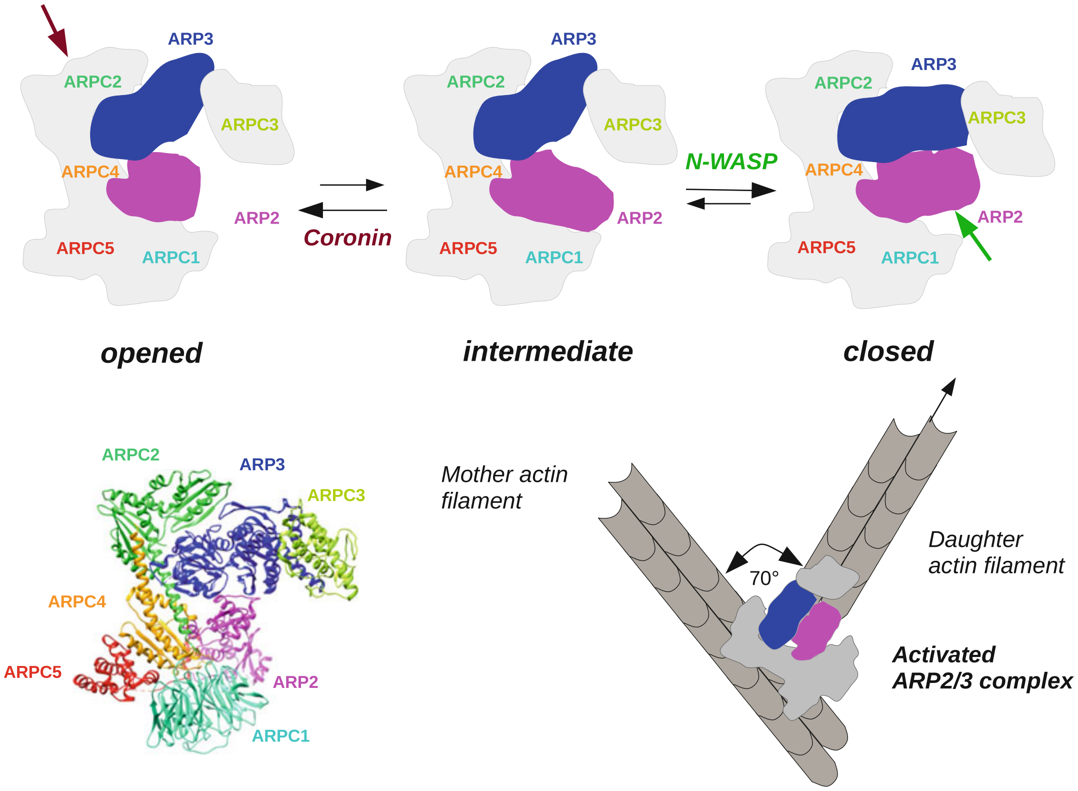


Fig. 20.4 **Structural alteration steps leading to active ARP2/3 complex**. Model of structural transition between inactive (opened), intermediate and active (closed) ARP2/3 complex observed in cryo-EM structural reconstructions. Terminal stages can be stabilized by inhibitor (coronin) or activator (N-WASP) proteins (their binding site is shown with *arrows*). Crystal structure of inactive ARP2/3 (PDB ID: 2P9L) fits the model of the opened state [36]. The figure on the right shows the branched actin filament bound by the ARP2/3 complex in its closed state. ARP2 and ARP3 (*magenta* and *blue*, respectively) mimic two actin monomers in the closed state of the ARP2/3 complex, thus they act as a template for the new filament growing in approximately 70° compared to the mother filament

20.6 Protein Complexes Controlling Transcription

A signaling pathway most often influences the transcription of selected genes. To understand transcription regulation, researchers in the last decade have reconstituted core transcriptional complexes [37]. These studies highlighted the importance of transient structural changes forming in response to activator or repressor molecules. The most studied complex is the class II transcription pre-initiation complex (PIC), which is a 4 MDa multi-protein assembly comprised of 60 polypeptides. PIC is comprised of RNA polymerase II (Pol II), general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH), and the Mediator complex [38] (Fig. 20.5). There are crystal structures available for some of the individual proteins and these could be used for docking them into the cryo-EM maps obtained on larger assemblies [39].

The 26-subunit Mediator complex acts like a bridge for signal transduction between transcription factors and RNA polymerase II. Its large surface area enables it to accept multiple inputs from transcription factors, co-activators, co-repressors, or nucleic acids. A simple input signal may be for example the appearance of an activated transcription factor on the DNA enhancer. According to the multiple allosteric network model, the input signal causes binding factor specific structural shifts which spreads across the whole complex

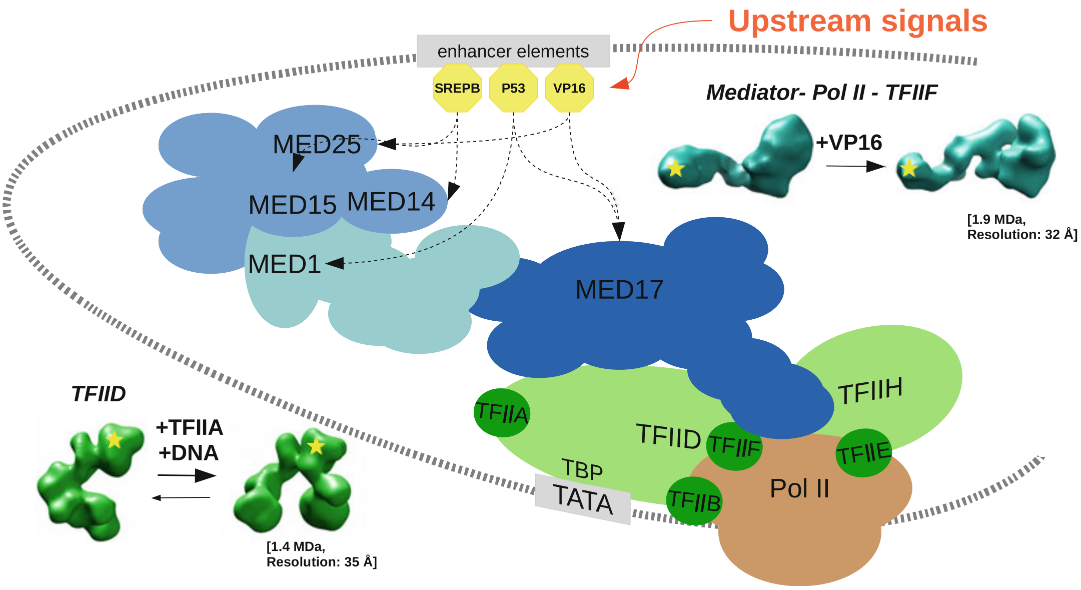


Fig. 20.5 Allosteric regulation in the transcriptional machinery. Schematic figure of the human PIC. Mediator complex (*blue particles*) has multiple binding sites for transcription factors (*e.g.*, p53, SREB and VP16). The binding of transcription factors or DNA to PIC subunits may cause structural shifts which leads to specifically regulated transcription. For example binding of VP16 transcription factor results in a structural shift in Mediator-Pol II-TFIIF assembly (*blue* cryo-EM maps; EMD-5344, EMD-5343 [52]. Cryo-EM maps (*green*) indicate directed reorganization of TFIID (EMD-2287, EMD-2284, respectively). The TFIID complex may exist in two distinct conformations, and binding of promoter DNA (TATA) and TFIIA stabilizes one of the conformations, which is competent to recruit Pol II [45]. *Yellow stars* indicate corresponding regions of the cryo-EM maps (*TBP* TATA binding protein)

[40]. This model suggests that the Mediator is best described not only as a loose network of interacting proteins but rather as a sophisticated multi-subunit complex with a network of different allosteric states. This mechanism helps to generate promoter specific outcomes through the PIC, which is comprised of ubiquitous components [41, 42]. Structurally explored examples are the sterol regulatory element binding protein (SREBP), p53 or the viral VP16 transcription factors that cause distinct structural shifts in the Mediator (Fig. 20.5). For p53, two of its domains may interact with two Mediator subunits, but interestingly only one binding mode brings about conformational changes that are compatible with Pol II elongation. The mechanism of Pol II activation is started with the binding of p53 activation domain to MED17, which in turn promotes TFIIH-dependent Pol II phosphorylation. Ultimately, the transcription machinery is now brought into its elongation competent state and transcription will start [43].

TFIID is also part of the PIC and it is composed of TATA-box-binding protein (TBP) and 13 TBP-associated factors (TAFs). Based on cryo-EM analysis the core-TFIID consists of two symmetric copies of TAF4, TAF5, TAF6, TAF9 and TAF12. In response to upstream signals, the TAF8–TAF10 complex is imported into the nucleus by importins, binds to the core-TFIID and breaks its symmetry. This results in an asymmetric 7TAF complex with new binding surfaces for six more TAF subunits and for TBP (canonical form) [44]. The promoter DNA and TFIIA trigger further structural changes and participate in the stabilization of the rearranged holo-TFIID complex (Fig. 20.5). The formation of the rearranged TFIID-TFIIA-DNA complex is then followed by binding of TFIIB, Pol II, TFIIF, TFIIE, and TFIIH to yield the transcriptionally competent pre-initiation complex [45].

For large multi-subunit complexes, single-particle electron microscopy (EM) is an essential method, especially when the sample is available in very little amounts. This technique combined with atomic resolution structures on monomers can give pseudo-atomic models on large complexes [46].

20.7 Conclusion

There are myriads of signaling complexes in action when cells respond to their environment. Fortunately, binary protein-protein interaction data on the proteome level is rapidly increasing thanks to systematic, large-scale protein-protein interaction studies and databases [47]. To what extent this wealth of information can be harnessed for mechanistic understanding of signaling complexes greatly depends on the reconstruction of functionally important signalosomes for structural analysis.

Obtaining atomic resolution structural information about a signaling question will require the reduction of a bigger complex into biochemically well-behaving smaller units, which have less disordered regions and are conformationally less heterogeneous. In this case the pitfall could be that higher-level biochemical properties of the whole complex may be lost in a reduced system. Fortunately, these smaller complexes could be built into low-resolution maps of bigger complexes. Ultimately, signaling complexes may be visualized *in cellulo* by super resolution microscopy techniques that are capable of breaching the 250 nm light diffraction limit by an order of magnitude [48]. This potentially bridges the resolution gap between structural reconstructions by X-ray crystallography/single-particle cryo-EM/cryo-electron tomography and the visualization of fluorescently labeled protein complexes via classical light microscopy in cells.

Acknowledgments This work was supported by a “Lendület” Grant from the Hungarian Academy of Sciences (LP2013-57) and by the Hungarian Research Fund (OTKA NN 114309).

References

1. Kobilka BK (2007) G protein coupled receptor structure and activation. Biochim Biophys Acta 1768(4):794–807
2. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. Science 289(5480):739–745
3. Yin X, Xu H, Hanson M, Liu W (2014) GPCR crystallization using lipidic cubic phase technique. Curr Pharm Biotechnol 15(10):971–979
4. Rasmussen SG, Choi HJ, Rosenbaum DM, Kobilka TS, Thian FS, Edwards PC, Burghammer M, Ratnala VR, Sanishvili R, Fischetti RF, Schertler GF, Weis WI, Kobilka BK (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature 450(7168):383–387
5. Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Yao XJ, Weis WI, Stevens RC, Kobilka BK (2007) GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. Science 318(5854):1266–1273
6. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi H-J, Kuhn P, Weis WI, Kobilka BK, Stevens RC (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 318(5854):1258–1265
7. Rasmussen SG, Choi HJ, Fung JJ, Pardon E, Casarosa P, Chae PS, Devree BT, Rosenbaum DM, Thian FS, Kobilka TS, Schnapp A, Konetzki I, Sunahara RK, Gellman SH, Pautsch A, Steyaert J, Weis WI, Kobilka BK (2011) Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. Nature 469(7329):175–180
8. Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah ST, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skiniotis G, Weis WI, Sunahara RK, Kobilka BK (2011) Crystal structure of the beta2 adrenergic receptor-Gs protein complex. Nature 477(7366):549–555
9. Zaidel-Bar R, Itzkovitz S, Ma’ayan A, Iyengar R, Geiger B (2007) Functional atlas of the integrin adhesome. Nat Cell Biol 9(8):858–867
10. Deakin NO, Pignatelli J, Turner CE (2012) Diverse roles for the paxillin family of proteins in cancer. Genes Cancer 3(5–6):362–370
11. Deramaudt TB, Dujardin D, Noulet F, Martin S, Vauchelles R, Takeda K, Ronde P (2014) Altering FAK-paxillin interactions reduces adhesion, migration and invasion processes. PLoS ONE 9(3):e92059
12. Wang Y, Chien S (2007) Analysis of integrin signaling by fluorescence resonance energy transfer. Methods Enzymol 426:177–201
13. Patla I, Volberg T, Elad N, Hirschfeld-Warneken V, Grashoff C, Fassler R, Spatz JP, Geiger B, Medalia O (2010) Dissecting the molecular architecture of integrin adhesion sites by cryo-electron tomography. Nat Cell Biol 12(9):909–915
14. Calderwood DA, Zent R, Grant R, Rees DJ, Hynes RO, Ginsberg MH (1999) The Talin head domain binds to integrin beta subunit cytoplasmic tails and

regulates integrin activation. J Biol Chem 274(40):28071–28074

15. Wegener KL, Campbell ID (2008) Transmembrane and cytoplasmic domains in integrin activation and protein-protein interactions (review). Mol Membr Biol 25(5):376–387
16. Legate KR, Fassler R (2009) Mechanisms that regulate adaptor binding to beta-integrin cytoplasmic tails. J Cell Sci 122(Pt 2):187–198
17. Schaller MD (2010) Cellular functions of FAK kinases: insight into molecular mechanisms and novel functions. J Cell Sci 123(Pt 7):1007–1013
18. Bertolucci CM, Guibao CD, Zheng J (2005) Structural features of the focal adhesion kinase-paxillin complex give insight into the dynamics of focal adhesion assembly. Protein Sci 14(3):644–652
19. Ginsberg MH, Partridge A, Shattil SJ (2005) Integrin regulation. Curr Opin Cell Biol 17(5):509–516
20. Bhattacharyya RP, Remenyi A, Yeh BJ, Lim WA (2006) Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev Biochem 75:655–680
21. Good MC, Zalatan JG, Lim WA (2011) Scaffold proteins: hubs for controlling the flow of cellular information. Science 332(6030):680–686
22. Bardwell L (2004) A walk-through of the yeast mating pheromone response pathway. Peptides 25(9):1465–1476
23. Good M, Tang G, Singleton J, Remenyi A, Lim WA (2009) The Ste5 scaffold directs mating signaling by catalytically unlocking the Fus3 MAP kinase for activation. Cell 136(6):1085–1097
24. Zalatan JG, Coyle SM, Rajan S, Sidhu SS, Lim WA (2012) Conformational control of the Ste5 scaffold protein insulates against MAP kinase misactivation. Science 337(6099):1218–1222
25. Dhanasekaran DN, Kashef K, Lee CM, Xu H, Reddy EP (2007) Scaffold proteins of MAP-kinase modules. Oncogene 26(22):3185–3202
26. Brennan DF, Dar AC, Hertz NT, Chao WCH, Burlingame AL, Shokat KM, Barford D (2011) A Raf-induced allosteric transition of KSR stimulates phosphorylation of MEK. Nature 472(7343):366–369
27. Dos Remedios CG, Chhabra D, Kekic M, Dedova IV, Tsubakihara M, Berry DA, Nosworthy NJ (2003) Actin binding proteins: regulation of cytoskeletal microfilaments. Physiol Rev 83(2):433–473
28. Goley ED, Welch MD (2006) The ARP2/3 complex: an actin nucleator comes of age. Nat Rev Mol Cell Biol 7(10):713–726
29. Rohatgi R, Ma L, Miki H, Lopez M, Kirchhausen T, Takenawa T, Kirschner MW (1999) The interaction between N-WASP and the Arp2/3 complex links Cdc42-dependent signals to actin assembly. Cell 97(2):221–231
30. Volkmann N, Amann KJ, Stoilova-McPhie S, Egile C, Winter DC, Hazelwood L, Heuser JE, Li R, Pollard TD, Hanein D (2001) Structure of Arp2/3 complex in its activated state and in actin filament branch junctions. Science 293(5539):2456–2459
31. Gournier H, Goley ED, Niederstrasser H, Trinh T, Welch MD (2001) Reconstitution of human Arp2/3 complex reveals critical roles of individual subunits in complex structure and activity. Mol Cell 8(5):1041–1052
32. Robinson RC, Turbedsky K, Kaiser DA, Marchand JB, Higgs HN, Choe S, Pollard TD (2001) Crystal structure of Arp2/3 complex. Science 294(5547):1679–1684
33. Beltzner CC, Pollard TD (2004) Identification of functionally important residues of Arp2/3 complex by analysis of homology models from diverse species. J Mol Biol 336(2):551–565
34. Goley ED, Rodenbusch SE, Martin AC, Welch MD (2004) Critical conformational changes in the Arp2/3 complex are induced by nucleotide and nucleation promoting factor. Mol Cell 16(2):269–279
35. Egile C, Rouiller I, Xu XP, Volkmann N, Li R, Hanein D (2005) Mechanism of filament nucleation and branch stability revealed by the structure of the Arp2/3 complex at actin branch junctions. PLoS Biol 3(11):e383
36. Rodal AA, Sokolova O, Robins DB, Daugherty KM, Hippenmeyer S, Riezman H, Grigorieff N, Goode BL (2005) Conformational changes in the Arp2/3 complex leading to actin nucleation. Nat Struct Mol Biol 12(1):26–31
37. Berger I, Blanco AG, Boelens R, Cavarelli J, Coll M, Folkers GE, Nie Y, Pogenberg V, Schultz P, Wilmanns M, Moras D, Poterszman A (2011) Structural insights into transcription complexes. J Struct Biol 175(2):135–146
38. Kandiah E, Trowitzsch S, Gupta K, Haffke M, Berger I (2014) More pieces to the puzzle: recent structural insights into class II transcription initiation. Curr Opin Struct Biol 24:91–97
39. Tsai K-L, Tomomori-Sato C, Sato S, Conaway RC, Conaway JW, Asturias FJ (2014) Subunit architecture and functional modular rearrangements of the transcriptional mediator complex. Cell 157(6):1430–1444
40. Lewis BA (2010) Understanding large multiprotein complexes: applying a multiple allosteric networks model to explain the function of the Mediator transcription complex. J Cell Sci 123(Pt 2):159–163
41. Poss ZC, Ebmeier CC, Taatjes DJ (2013) The Mediator complex and transcription regulation. Crit Rev Biochem Mol Biol 48(6):575–608
42. Carlsten JO, Zhu X, Gustafsson CM (2013) The multitalented Mediator complex. Trends Biochem Sci 38(11):531–537

43. Meyer KD, Lin S-C, Bernecky C, Gao Y, Taatjes DJ (2010) p53 activates transcription by directing structural shifts in Mediator. Nat Struct Mol Biol 17(6):753–760
44. Bieniossek C, Papai G, Schaffitzel C, Garzoni F, Chaillet M, Scheer E, Papadopoulos P, Tora L, Schultz P, Berger I (2013) The architecture of human general transcription factor TFIID core complex. Nature 493(7434):699–702
45. Cianfrocco MA, Kassavetis GA, Grob P, Fang J, Juven-Gershon T, Kadonaga JT, Nogales E (2013) Human TFIID binds to core promoter DNA in a reorganized structural state. Cell 152(1–2):120–131
46. He Y, Fang J, Taatjes DJ, Nogales E (2013) Structural visualization of key steps in human transcription initiation. Nature 495(7442):481–486
47. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41(Database issue):D808–D815
48. Sahl SJ, Moerner WE (2013) Super-resolution fluorescence imaging with single molecules. Curr Opin Struct Biol 23(5):778–787
49. Tesmer JJ, Sunahara RK, Gilman AG, Sprang SR (1997) Crystal structure of the catalytic domains of adenylyl cyclase in a complex with Gsalpha. GTPgammaS. Science 278(5345):1907–1916
50. McCudden CR, Hains MD, Kimple RJ, Siderovski DP, Willard FS (2005) G-protein signaling: back to the future. Cell Mol Life Sci 62(5):551–577
51. Ikeda SR (1996) Voltage-dependent modulation of N-type calcium channels by G-protein beta gamma subunits. Nature 380(6571):255–258
52. Bernecky C, Taatjes DJ (2012) Activator-mediator binding stabilizes RNA polymerase II orientation within the human mediator-RNA polymerase II-TFIIF assembly. J Mol Biol 417(5):387–394

Part VII

Biophysical Methods to Assess Protein Complexes

The Use of Small-Angle Scattering for the Characterization of Multi Subunit Complexes

21

Adam Round

Abstract

As the continuing trend in structural biology is to probe ever more complex systems, new methodologies are being developed plus existing techniques are being expanded and adapted, to keep up with the demands of the research community. To investigate multi subunit complexes (protein-DNA, protein-RNA or protein-protein complexes) no one technique holds a monopoly, as each technique yields independent information inaccessible to the other methods, but can be used together in a complementary way. Additionally as large conformational changes are not unlikely, investigation of the dynamics of these systems under physiological conditions is needed to fully understand their function. Investigations under physiological conditions in solution are becoming more standardized and with more dedicated, automated beamlines available these experiments are easy to access by the general research community. As such the need for explanations of how to plan and undertake these experiments is needed. In this chapter we will cover the requirements of these experiments as well and how to plan undertake and analyze the results of such experiments.

Keywords

SAS • X-rays • Neutrons • Contrast • Solution scattering • SEC • BioSAS • Complementary techniques

A. Round (✉)
European Molecular Biology Laboratory,
Grenoble, 71 avenue des Martyrs, 38042, France

Unit for Virus Host-Cell Interactions, University
Grenoble Alpes-EMBL-CNRS,
71 avenue des Martyrs, 38042 Grenoble, France

Joint Structural Biology Group, European
Synchrotron Radiation Facility,
Grenoble, 71 avenue des Martyrs, 38042, France

Faculty of Natural Sciences, Keele University,
Keele, Staffordshire ST5 5BG, UK
e-mail: around@embl.fr

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_21

21.1 Introduction

To investigate multi subunit complexes (protein-DNA, protein-RNA or protein-protein complexes) no one technique holds a monopoly. NMR is most feasible for smaller proteins <35 kDa [1] and although with labeling strategies this upper limit can be extended [2] currently less than 2 % of the NMR structures deposited in the protein data bank are over 30 kDa. EM which has traditionally been limited to Mega-Dalton sized complexes is now not only approaching atomic resolution [3, 4] but achieving this with smaller complexes [5], with the current minimum being 170 kDa [6]. Crystallographic studies which have no intrinsic size restriction fill the gap between NMR and EM. However, as its name suggests X-ray crystallography is absolutely dependent on obtaining crystals of the macromolecule (or complex) [7]. As the tendency to crystallize is reduced for systems with conformational flexibility [8], it can be difficult if not impossible to obtain for some systems. If you are lucky enough to obtain crystals and solve the structure of your complex, flexible and dynamic regions can be fixed in a specific conformation because of crystal packing interactions, which are artifacts and play no part in function [9]. The influence of crystal packing on atomic fluctuations is very important [10], as a consequence is the conformations observed in crystal structures may not actually be an accurate representation of the structure in solution [11]. As large conformational changes are not unlikely in multi domain complexes, investigation of the dynamics of these systems under physiological conditions is needed to fully understand their function. Scattering experiments of biological macromolecules in solution using both neutrons and X-rays is ideal (especially when combined with information from other techniques) for such investigations as the components of the solution can be adapted to mimic physiological conditions and varied experimentally to enable functional investigation [12].

Scattering of X-rays and neutrons are highly complementary and can be mathematically described together even though their interaction with the matter is different. X-rays are scattered by charged particles (predominantly electrons) whereas neutrons interact with the nuclei (its nuclear potential and spin). This fundamental difference gives rise to main difference in the techniques as the contrast of X-rays is proportional to the electron density and thus to the atomic number of the element, whereas the neutron contrast is independent of atomic number and can be dramatically different between isotopes of the same element. The most notable difference between hydrogen and deuterium is used for contrast variation in neutron scattering by adjusting the ratio of water (H_2O) and "heavy" water (D_2O) in the surrounding buffer. The measured intensity is proportional to the contrast (difference between scattering power) of the scatter (protein) and its environment (buffer). As this contrast is small for both X-rays and neutrons strong incident intensity is required to provide a measurable signal and thus performing such experiments at large scale facilities (synchrotrons, reactors, spallation sources) is preferable.

Scattering experiments observe the intensity of scattered radiation at a range of angles to the incident beam. Depending on the distance from separated scattering centers and the observation point, there is a resulting phase shift and thus an observable interference pattern. Given a fixed (monochromatic) incident beam (typically $\lambda = 0.1$–0.15 nm and 0.1–1.0 nm for X-rays and neutrons respectively) to observe length scales typical for proteins (0.5–50 nm) the angular deviations of interest are small (few degrees) hence such experiments are referred to as small angle scattering (SAS).

SAS experiments of biological macromolecules in solution (bioSAS) using both neutrons (SANS) or X-rays (SAXS) provide data on the size and shape of the scattering object. Using Guinier's law [13] the radius of gyration (Rg), a measure of the overall size, can be determined and the intensity at zero angle (I_0), which is proportional to molecular mass (when scaled for the number of scatterers, *i.e.*, protein concentration), can be extrapolated. Additionally the hydrated volume of the scatterer can be determined using Porod's law [14] and the maximum dimension (D_{max}) estimated through the process of the inverse Fourier transform [15, 16]. Unfortunately as proteins in solution are mobile, all orientations are possible. Combined with the intrinsic lack of phase information (only the intensity can be measured) any shape reconstructions (although a powerful tool), are by nature ambiguous. Furthermore, as such modeling programs impose the intrinsic restriction that an individual model is created which fits the data, the assumption (of monodispersity) that the data is from an average of identical scatterers, which scatter independently (only form factor) must be validated. SAS is not a new technique, the first experiments date back to the 1930s. However, in recent years the combination of advances in sample production, high powered (X-ray and neutron) sources with rapid access to automated systems and advanced modeling (taking advantage of modern computing) has made bioSAS a viable and desirable tool for the structural biologist. In order to aid those wishing to exploit BioSAS this chapter covers the requirements of these experiments as well as how to plan for, undertake and analyze the results of such experiments.

21.2 Requirements

Although it is not possible to give comprehensive instructions with precise values needed for a SAS experiment with multiple subunits and their complexes, as it will depend on where the experiment will be done and on the project itself, the following information should be a guide to help understand what the important considerations are when planning. Once the outline of the experiment is defined, consultation with the staff of the facility you hope to use will be necessary to clarify the logistics such as when its possible shipping details, how much sample is required and how long the data collection should take.

21.2.1 Sample Preparation

As with any individual SAS experiment each sample is required to be not only to be pure but monodisperse, *i.e.*, a single oligomeric species in the same conformation. Thus before considering a SAS experiment thorough sample preparation and cross checking is necessary.

21.2.1.1 Sub-complexes and Individual Components

Additional consideration must be given to experiments for multi subunit complexes as a thorough experiment requires a comprehensive data set which includes not only the complex of interest but all the individual sub components and any partial constructs which can be assembled. This may also include truncated forms of the subunits if of interest. The more parts of the complex, which can be measured the more cross checks and constraints, can be applied during data processing and analysis. Even if the structures of subunits are available it is still advisable to measure SAS data for them to verify their behavior in solution.

21.2.2 Quality Control and Checks Priors to SAS

For every construct that is to be measured, the monodispersity should be confirmed with any and all techniques available. Additional information such as molecular mass (MM), binding behavior and stoichiometry from any available technique is also valuable during data analysis. In short all available techniques which aid your understanding of the sample, its behavior and thereby increase confidence that your samples are suitable for SAS experiments can and should be used wherever possible.

Analytical gel filtration is a commonly used technique for the final purification of samples and as such is generally viewed and a standard requirement prior to SAS studies (more details below). Light scattering can be very useful for assessing the monodispersity of samples and providing additional information on the their size, dynamic light scattering is available in many laboratories as standard whereas MALLS (which is often also combined with SEC) which may be preferred is not as widely available but should be used if possible. Additionally for a complexes, their binding should be investigated, Analytical Ultracentrifugation (AUC) is therefore a powerful technique to use as it provides information on binding and stoichiometry as well as provides additional checks on monodispersity, MM and shape (globular or unfolded). However, as AUC is not readily accessible ITC is more commonly used to assess binding.

21.2.2.1 Buffer Optimization (One for Each, Not One for All)

An unfortunate complication of multi subunit complexes especially those including nucleic acids and protein is that the buffer conditions that are best for the nucleic acids are not necessarily the same as those for the protein and can be yet again different for the complex itself, therefore optimization of the buffer for each subunit is necessary. This can naturally lead to the question of the behavior of the parts with the changing conditions and so additional measurements could be necessary to understand the effects of conditions on the subunits in order to be able to exclude those effects from the conclusions.

21.2.2.2 Purification and Dissociation

Another check that should be done is regarding the assumption of purity following purification with gel filtration. Although the fraction collected will be pure complex there is no guarantee that the sample will be stable. Even with a strong binding constant the sample might dissociate over time and so the purified fraction should be concentrated and the gel filtration repeated to test its behavior over time. Additionally, the effects of freezing and thawing should also be tested by repeating all the verification steps after freeze thaw cycles. If there is significant dissociation with time or freeze thaw cycles, preparation or perhaps a final purification of the complex immediately prior to the data collection should be considered.

21.3 Designing Your Experiment

Practical experimental considerations in terms of time for data acquisition and volume per measurement will to a large extent be governed by the facility used for the experiment. Specific details for each instrument are published but in general for X-ray facilities you will need 10–100 μL per sample and measure for seconds whereas for Neutron sources you may need 100–400 μL and will measure for minutes to hours.

The most important consideration for the samples is not only the complexity of the system (number of subunits and how they interact) but also the behavior of the samples (including all the individual domains and partial subunits as well as the full complex).

For all structural studies using biological samples much of the time and effort for the project is for sample production. Cloning, expression, purification, and optimizing the conditions can be time consuming, and yields are nearly impossible to predict as they vary greatly between systems. Once the purified samples have been obtained a complete biophysical characterization should still be required to ensure that the sample is in its active state. Structural studies can then proceed using the pure sample. However, further sample preparation may be needed.

Each technique has its benefits and no one structural technique gives all information; as such, combining techniques is increasingly used to tackle complex biological problems. Table 21.1 below gives a brief overview of the time and sample requirements for different techniques for comparison.

21.3.1 Practical Considerations: Current State of the Art

Sample volume is always a consideration and one that facilities are working continually to reduce. However, volume is not only a matter of the size of the sample holder but also of the behavior of the sample. The more sensitive the sample is to radiation (further details in Sect. 21.4.1) the more sample will be required. Each facility will provide recommendations for required sample volume of all the data collection options provided which will help inform you of the material required and thus the time required for sample preparation prior to the data collection.

Table 21.1 Values presented here are for a single sample and are a guide based on expected usage and can vary depending on complexity of project and sample behaviour

Technique	Sample required	Preparation	Experiment	Analysis to 3D	Comments
MX	1–10 mg	Months to years	Minutes	Days	High-resolution structure obtained if highly diffracting crystals can be produced
	Dependent on success of crystallization	For crystallization, screening and optimization		Provided phases can be solved	Crystal packing can result in modifications to the structure
EM	0.01–0.1 mg	Days	Hours	Months	Medium resolution can be obtained in cases with symmetry
					Limited to samples bigger than 300 kDa
					Sample preparation can result in artifacts
NMR	5–15 mg	No additional preparation beyond purification	Hours	Months	High resolution possible but limited to proteins smaller than 25 kDa
SAXS	0.1–0.5 mg	No additional preparation beyond purification	Minutes	Days	Low-resolution data obtained under near physiological conditions. No intrinsic size limits
SANS	0.5–2 mg	As for SAXS but additional days to weeks for labeling if required	Hours	Days	As for SAXS but with the addition of contrast additional on domain positions can be obtained

21.3.1.1 Static Data Collection with Current High Throughput Systems

Many synchrotron SAXS beamlines, especially those specializing in biological samples in solution, *i.e.*, EMBL-HH [17, 18], Soleil [19], ALS [20], and ESRF [21, 22] (Fig. 21.1) are now equipped with automatic sample changers. The overriding aim of such systems is to facilitate measurements of many samples under different conditions with confidence. Sample volumes in the range of 30–50 μL per measurement can be used as standard, which, for a typical dilution series of three concentrations, should require maximally 100 μL of the highest sample concentration (with volume left for additional measurements if required). Although there is variation between facilities and samples, standard data acquisition protocols using state-of-the-art sample changers allow for full data collection of a complete concentration series (with a minimum of three samples) including background buffer measurements (before and after each sample), with thorough cleaning and drying between all measurements, in a little over 5 min.

21.3.1.2 Online Size Exclusion Chromatography

Additional measurement strategies are also available, in particular, online size exclusion (SEC) chromatography (both FPLC and HPLC) is becoming increasingly popular [19, 23] as it enables separation of samples which are inherently mixtures as they show association and dissociation resulting in an equilibrium mixture and therefore even if it is purified the sample will still be a mixture when it is measured. The use of SEC allows separation of species from dynamic mixtures allowing analysis of the individual components and the behavior of the mixture under physiological conditions, which may be biologically relevant. Sample volume required,

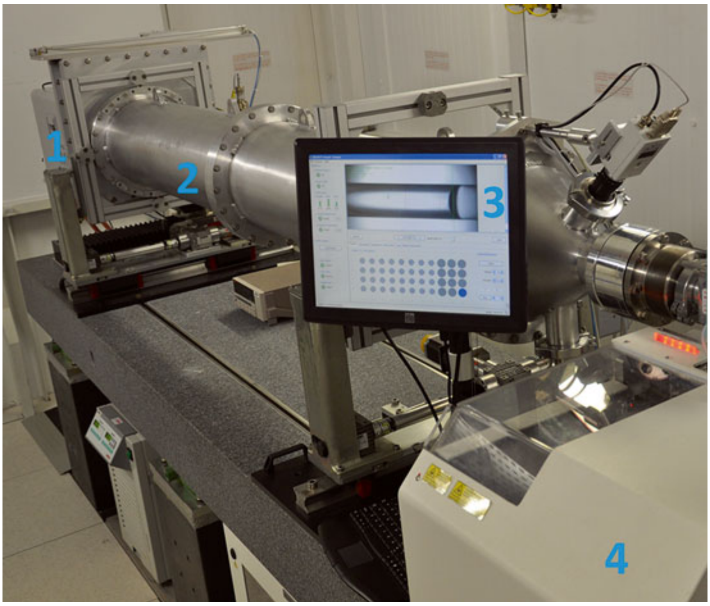


Fig. 21.1 Experimental setup of the ESRF BioSAXS beamline BM29. This experimental facility is dedicated to SAXS measurements of samples in solution offering both static (batch) operation and online SEC measurements (both HPLC and FPLC). X-ray scattering images are acquired using a Pilatus 1M detector (*1*). Air scattering is avoided by using an evacuated flight tube (*2*). A touch screen monitor (*3*) allows easy control of the dedicated sample changer (*4*)

especially with modern (small volume) columns, is similar to static measurements (50–100 μL total) with an injection possible every 10 min. However, actual sample consumption is a function of the column used and data acquisition time a function of flow speed used. In general you should already have defined the SEC protocol to ensure adequate separation of your peaks (choice of column, volume and concentration to inject and flow speed) prior to the online SAXS data collection. The aim is to ensure maximum separation of all species present with the highest concentration possible (without saturating the column) to maximize signal to noise ratio. As the SEC protocol is defined to meet the needs of the sample, the variations between facilities are related more to the data acquisition parameters (number of frames and method for processing) than to the total time or material required, with the exception that you could decide to average the data acquisition over multiple injections to increase the signal to noise ratio.

21.3.1.3 Integrated Data Collection

As SEC data acquisition dilutes the samples, the concentration decreases and therefore the signal at wider angles can become quite noisy. By measuring the sample with online SEC and, also in static mode, it may be possible to combine both data sets to make the idealized curve, provided that the presence of artifacts in the static data can be excluded. The aim of the integrating static data acquisition with SEC is to maximize data quality and confidence that the resulting idealized curve is free from artifacts. The SEC data gives the low angle data with verification that the Rg is free from aggregates (as they will have been separated on the column) and the wider angle data from the high concentration static data which will have the best signal to noise ratio.

Many facilities offer both static high throughput and online SEC measurements. However, not all facilities currently provide easy switching between the two data collection modes. Some facilities, such as BM29 at the ESRF, have integrated systems (Fig. 21.2) that allow rapid exchange between the data collection methods. These systems allow users to control the scheduling to enable static and SEC measurements to be interspersed to provide maximal efficiency for data acquisition. In the absence of such systems switching between static and SEC measurements may require manual (local expert) intervention, depending on the specific setup and thus may only be possible during supported working hours.

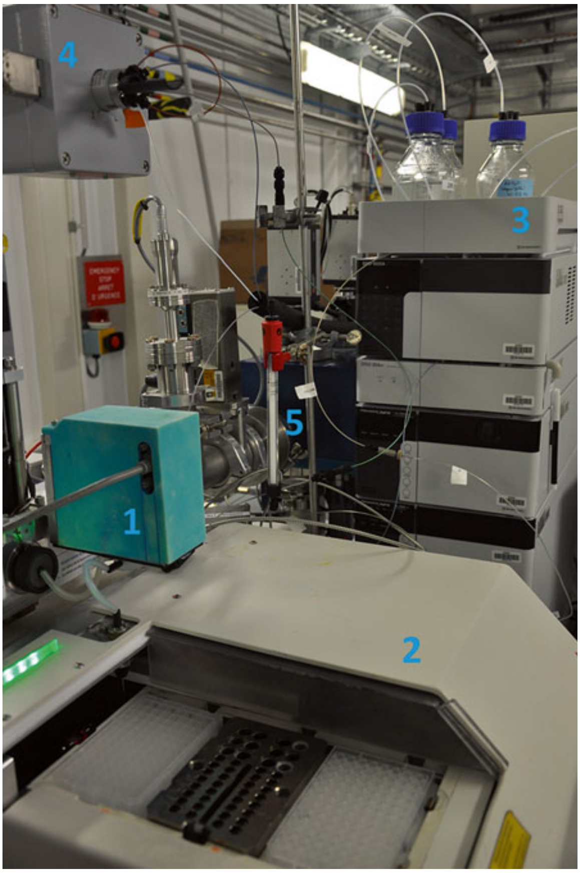


Fig. 21.2 Integrated static and SEC operation software-controlled valve (*1*) allows safe switching between sample changer (*2*) or SEC (*3*) operation. A second software-controlled valve (4) enables manual triggering of injection of sample onto column (*5*). Fast automated switching between static and SEC modes, allows the users to optimize the use of both modes to make the most efficient use of their beamtime

21.3.2 Combined Experiments

It should be noted that an experiment can and should utilize any and all options for data acquisition in combination (whenever possible) to maximize the information one can obtain. If there is access to a lab source and the samples are stable enough, it can be used in combination with synchrotron data.

The use of neutron scattering and contrast variation can provide additional information on samples such as protein-nucleic acid complexes and by using labeling strategies, this can also be used for protein-protein complexes and is especially useful in cases where there is a known conformational change on biding (see Sect. 21.5.5).

It is advantageous to have characterized your samples using as many techniques as possible (including) X-rays to support the application for neutron beamtime. Even if you have already measured SAXS from your samples previously is can be highly advantageous to repeat these measurements on the same samples used for the neutron experiment (ideally at the same time). Joint SAXS and SANS experiments are therefore recommended to make the best and most efficient use of the samples. As SAXS data will not be affected by D_2O content in the buffer by collecting SAXS data, on any (or all) of the deuterated (neutron contrast) data sets, you should observe the full complex. Additionally, if the SAXS data is the same at all contrast match points measured (typical values are 0, 40, 75 and 100 % D_2O), the SAXS data provides verification that the variation seen using SANS is solely due to the contrast and not aggregation caused by deuterated buffers.

21.4 Data Acquisition

Once the initial experimental conditions have been defined the experiment can proceed. However, the initial conditions are subject to modification based on the feedback from the experiment itself; each condition measured should be checked to ensure it is of maximal data quality and each construct should be crosschecked to ensure its behavior is understood and the idealized curve can be made.

21.4.1 Considerations to Maximize Data Quality

There are unfortunately no ideal combinations of conditions that will give the perfect data, as with most experimental practice all parameters are a compromise. Higher concentration means better signal, but also increased sample requirements and increased likelihood of interparticle scattering. As each construct behaves differently, it's behavior when measured might indicate the need for additional measurements and crosschecks, which needs to be anticipated during the planning stages. However, understanding what compromises will have to be considered will enable better planning and decision-making and will therefore maximize the likelihood that the best possible data can be obtained from the available samples.

21.4.1.1 Additives: Signal to Noise Versus Sample Stability

As discussed previously buffer optimization is needed for each construct and will vary, depending on the surface charges pH might be altered, salt concentration will need to be altered, glycerol might be needed. For some experiments additional additives/ligands might be needed to ensure the construct in is its physiologically relevant state. Any additional compounds present in the buffer will affect the scattering to a greater or lesser degree. Anything present in the solution will attenuate the X-rays, it does not matter if this attenuation is of the direct beam or the scattered radiation the effect on the data will be the same. There is scattering from the additives to be considered but for small molecules this contribution will not be significant at the length scales observed in a SAS experiment. Special consideration should be given to additives such as glycerol as this in addition to the attenuation it causes also increases the electron density of the buffer thereby reducing the contrast and the scattering intensity from the sample.

General advice for any additives added is to ensure they are at the same concentration in the background buffer; unfortunately, this can be difficult if the additive is interacting with the sample. Additional measurements of the additive at higher concentration can also be considered to verify at excess what effects the additive can cause on the scattering so these effects can be taken into account in analysis or ideally its effect ignored as insignificant.

21.4.1.2 Lipids

Special consideration is needed for membrane proteins. As transmembrane domains are not generally soluble in aqueous solution without the addition of lipids. However, lipids are strong scatterers of both X-rays and Neutrons. Though neutron scattering offers the possibility to use contrast variation to match out the lipid component of the scattering, it is also useful to see the whole complex including the lipids in order to fully characterize the complex. When visualizing the lipid component care must be taken that there are no micelles formed. Special care should be taken during sample preparation as the micelles can be inadvertently concentrated along with the protein. Even by treating the sample and buffer in the same way this will not necessarily give an ideal matching background measurement and the strong signal of the lipid scattering will be detrimental to signal to noise. Therefore the most reliable method to give accurate subtractions is to utilize online size exclusion chromatography which is available it a number of synchrotron beamlines as standard (see Sect. 21.3.1). The choice of lipids is critical in order to ensure the micelles formed are of a significantly different size to enable maximal separation.

21.4.1.3 Important Note for Additives and Sample Quality

Whatever conditions are necessary to ensure the sample is in the correct state cannot be avoided. There is no benefit to measuring a stronger signal from a sample which is not is the state of biological interest. However, care should be taken that additives are not in excess, unless you can show their effect is minimal. In the cases where there is a noticeable effect on data quality it is preferable to accept that the counting statistics will be lower and to adapt the data acquisition to compensate where possible in order to ensure the measurement is of the sample in the correct state.

21.4.1.4 Radiation Damage

Absorption effects can be significant with X-ray scattering and are due principally to the photoelectric effect. For neutrons, absorption by nuclear capture (which can lead to decay events and damage) is usually negligible except for nuclei such as cadmium, boron, gadolinium, and other rare earth elements. These elements being rare means in practical terms absorption effects (and possible resulting complications) for neutrons are negligible. In Addition being electrically neutral, neutrons do not produce free radicals (as X-rays do) and therefore do not cause significant radiation damage [24]. Therefore even though neutron beams have a relatively low flux "damage" free data can be collected from a single aliquot with increased exposure time.

Interactions between X-rays and biological samples and their effects are well documented for crystalline samples [25]. It must be assumed that free radical production and bond brakeage will also be present in biological samples in solution. However, the resolution of solution scattering is not adequate to visualize these effects. What is commonly observed in bioSAXS experiments and termed "damage" is X-ray induced aggregation and/or precipitation (Fig. 21.3). This aggregation is caused by the secondary effects, *i.e.*, the action of the free radicals (which were produced by interaction of the X-rays and the sample) producing additional charged areas on the surface of the sample, which through electrostatic attraction causes aggregation.

Radiation effects are proportional to the absorbed dose. However, this is also related to the propensity to produce free radicals which is dependent upon the buffer and its additives more than the sample, as in solution scattering the sample is by design in a dilute state (often much less than 5 mg/mL). Increased salt concentration and the presence of any additives with heavy atoms will produce more free radicals for the same X-ray exposure. Thus avoiding an excess of salt and other additives is advisable.

The effect of the free radicals on the sample is highly dependent on sample properties, as the surface charges of the sample have a strong effect on the charge needed to promote aggregation. Multi-subunit complexes may therefore be strongly affected if the interaction between domains is electrostatic in nature. This is observed for protein-DNA complexes, which manifest a greater tendency to suffer radiation-induced aggregation than the protein alone [26]. Another significant factor regarding the effect of radiation damage on a sample is the stability of the sample in its buffer. For instance, lysozyme at pH 7.5 shows significantly more susceptibility to the effects of X-rays than at pH 4, as lysozyme is more stable under acidic conditions. This effect can be compounded in the case of multi-subunit complexes as each subunit may require specific buffer conditions for stability, which may again require optimization upon formation of subcomplexes and the full complex.

Although the primary effects of X-ray radiation are well known, using this knowledge to predict effects for solution experiments is non-trivial due to the lack of structural knowledge (surface charges) which is essential to hope to predict the behavior in solution. Although there is work towards understanding the effects of X-ray radiation on solution samples for bioSAXS experiments, currently an empirical approach is adopted to assess the effect of X-rays on biological samples.

The use of high frame rate pixel detectors is helping improve data quality. Their use is now makes it possible to routinely measure multiple frames for all samples and simultaneously monitor any variations from the initial frame (Fig. 21.3).

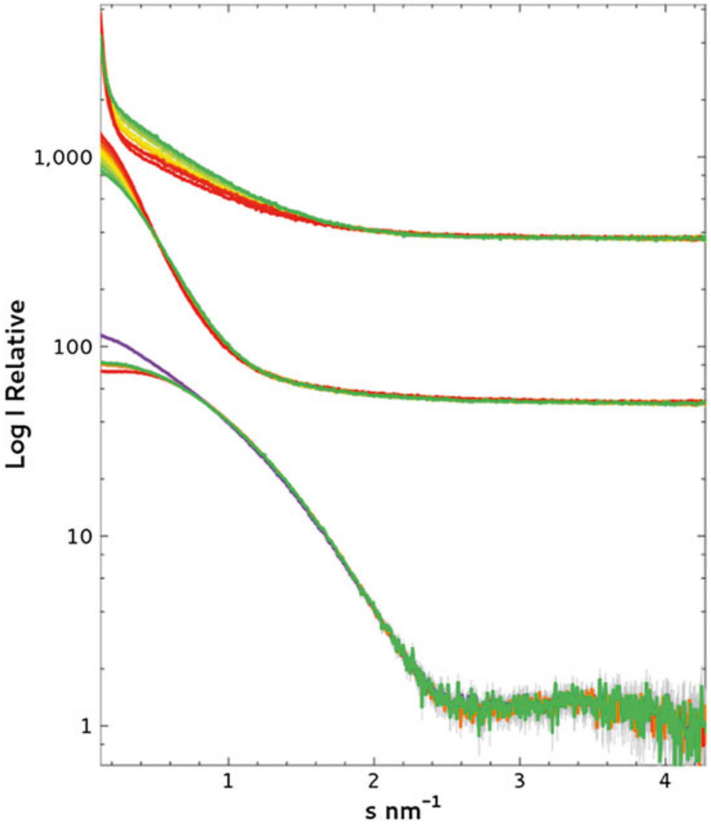


Fig. 21.3 Common problems in data acquisition. Precipitation (*top*, lysozyme at 5 mg/mL) and Aggregation (*middle*, BSA at 5 mg/mL) induced by exposure to X-rays. *Top* and *middle* plots show 25 individual 1-s frames acquired without flow at ESRF BM29 (*green*, first frame, to *red*, last frame). As the sample precipitates (scattering intensity is reduced) fewer scatterers are present in solution. As sample aggregates (intensity at low angles increases) there is an increasing contribution from larger scatterers in solution. Averaged (radiation damage free) data once subtracted (*bottom*, lysozyme) can often display effects with varying concentration, either repulsive, as shown by a decrease (*orange*, 5 mg/mL, and *red*, 10 mg/mL), or aggregation, as shown by an increase (*purple*, 10 mg/mL) of the scattering intensity at low angles compared to the lowest concentrations (*green*, 2.5 mg/mL). The nature of the interactions at higher concentrations are dependent on surface charges on the scatterers and so can vary with varying salt content and pH (*red*, pH 4, 200 mM NaCl; purple, pH 7.5, 0 mM NaCl)

Deviations from the initial frame might alert the user to the possibility that significant radiation damage has accumulated on the sample. Unfortunately, even in the absence of such warning signs it is not possible to state conclusively that there are no radiation effects, as primary radiation damage will always occur even at low dose. Secondary effects may also be occurring but their adverse effects may take longer to accumulate than the data acquisition time. Despite the caveats, it is generally assumed that if the sample shows no variation from the initial frame, effects of radiation can be considered to be minimal in terms of the low-resolution data.

21.4.1.5 Mitigating Strategies for Radiation Damage

X-ray induced effects on the sample may be unavoidable but the aggregation/precipitation which negatively affects bioSAXS experiments can be mitigated using a number of strategies. The choice of the most appropriate strategy to use (alone or in combination) will depend on the specifics of the project.

Table 21.2 List of common additives and their effects on data acquisition

Type		[a]Suggested concentration	Attenuation	Background	Radiation effects	Comments
Free radical scroungers	DTT dithiothreitol	mM			–	
	BME β-mercaptoethanol	mM			–	
	Ascorbate	mM			–	
Membrane mimicking	Lipids		+	+		Danger of micelles
	Detergents			+		Danger of micelles
Other	Salts	<500 mM	+		+	
	Heavy metals		++	+	+	
	Ligands		+ Size dependent	+	+	
	Glycerol	<5 % by volume	+	+		Reduces contrast between sample and buffer thus further reducing the observed scattering signal
Nucleotides	ATP, ADP, etc.	mM	+		Possible	

[a]Suggested concentrations are guidelines with regards to data quality (counting statistics). However, in some cases samples will need higher than this recommendation in order for the sample to be in the state desired for investigation (see Sect. 21.4.1)

The addition of free radical scroungers (see list of additives in Table 21.2) can be an effective strategy for many proteins, as they will limit the production of the charge centers and thus the propensity to aggregate. However, the effect of the scroungers on the sample itself should be assessed to ensure the data collected is not affected by their presence. If scroungers can be added without detriment to the sample then they can be used, though often freshly prepared aliquots will need to be added to ensure maximum efficacy and care should be taken to make sure the same concentration in the background buffer too. If an effective scrounger (which does not affect the sample) cannot be found then other mitigating strategies must be employed.

As radiation damage effects scale with absorbed dose (the energy deposited per unit mass), increasing the sample volume (for the same exposure) will therefore reduce the dose without affecting the data quality. However, this is at the expense of using greater amounts of sample, which may not be available in all cases.

In many experimental setups the volume of sample in the cross section of the beam position is small (less than 1 μL). However, the volume used for a measurement is generally greater (30, 50, 100 μL or more). This is due to the need for the sample holder to be larger than the beam but also so that the sample flows (or sometimes oscillates) through the sample holder during exposure in order to spread the dose over a larger volume, taking advantage of the assumption of sample homogeneity. This has the additional benefit that in flow mode you are constantly illuminating a fresh (un-irradiated) sample and slow secondary effects will not affect the data collected as the "damaged" sample volume will no longer be in the X-ray beam.

If additional sample mass is not available dilution of the sample can be effective, although data quality will also be affected as scattering intensity is proportional to sample concentration. Not only does this allow faster flow of the sample giving the benefits of spreading the dose and removing the "damaged" sample volume, but additionally as the sample is more dilute the particles are farther apart and so the electrostatic attraction should be reduced and the effects on aggregation reduced. In practice, however, it can be sometimes observed that the lowest concentration has the strongest effect of radiation-induced aggregation as the aggregation produced has a more significant effect on the average than at higher concentrations, where the unaffected volume dominates the scattering signal. Thus the concentration series, which should be collected as standard, is needed to assess these effects.

The flow speed can be manipulated to maximize data collection efficiency. Flow speed is a function of both the sample volume and the exposure time. Hence, should the volume available be unavoidably fixed, then reducing the exposure time will also increase the flow speed thereby giving the associated benefits, although data quality may be reduced as a result.

Reducing the X-ray flux will reduce the dose at sample position but will also reduce the scattered intensity proportionally. In some cases it has been claimed that there is a dose rate threshold below which the sample will survive longer, thereby allowing a gain in data quality by measuring longer. However, evidence for this threshold is limited and will probably be sample dependent, so it is not practical to exploit it routinely. However, if the other mitigating strategies cannot be employed for any reason then reducing the incident flux is a viable option.

21.5 Data Processing and Common Pitfalls

Data processing and analysis is not a separate part of the experiment which commences once the data acquisition is complete, it is an integral part of the data collection as preliminary results of data reduction and analysis provide valuable feedback on data quality.

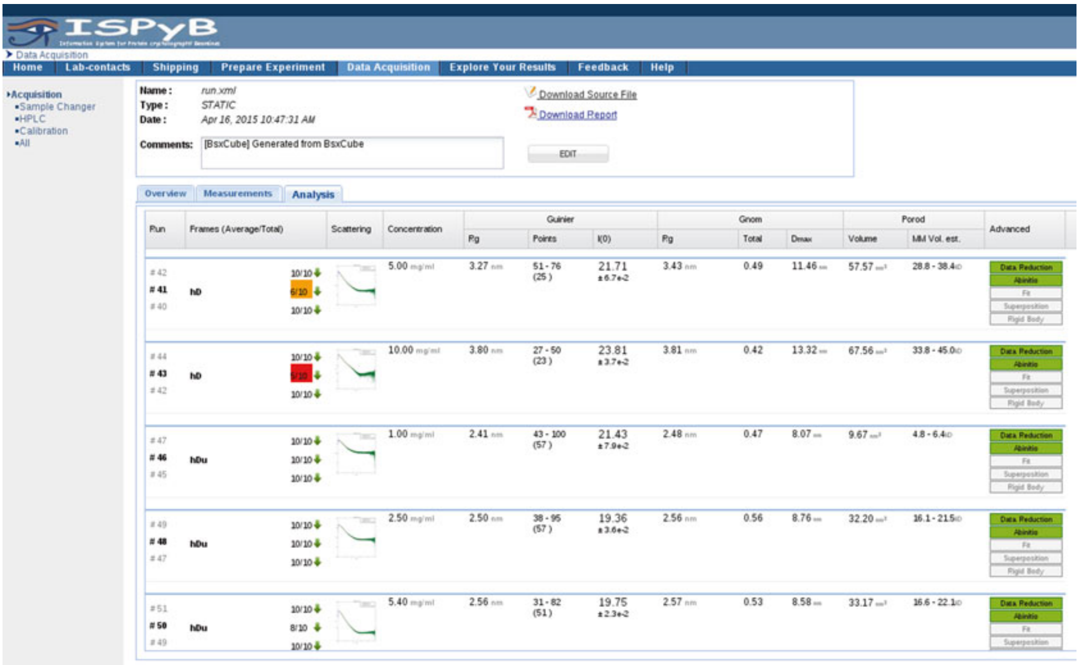
Many beamlines, which undertake bioSAXS experiments, have in recent years adopted an automated approach to data collection as well as preliminary processing of [21, 27]. These tools provide the useful invariants (Rg, I_0, Volume, MM estimates and D_{max}), which give valuable feedback regarding the sample behavior and data quality (Fig. 21.4).

There are a number of factors, which can affect data quality (nonspecific aggregation, radiation damage, flexibility, contamination and mixtures), and therefore the success of an experiment. For experiments on multi-subunit complexes the possibility of mixtures is increased, not only depending on the sample preparation procedures. Equimolar mixtures typically cannot be prepared with absolute accuracy and will invariably result in excess components. If the multi-subunit sample is purified as a complex, degradation and/or disassociation of some components might still occur. Furthermore, depending on how well the subunits bind, it may simply not be possible to have a fully bound complex, as there is continual association and disassociation. Unfortunately, at low resolution there can be similarities in the way the effects of these factors manifest and thus careful analysis of the data is required.

21.5.1 Concentration Effects

BioSAS experiments are almost exclusively interested in the form factor of scattering particles. However, what is accessible to experimental measurement is the combination of the form factor and the structure factor. The effect of the structure factor in SAS measurements can be minimized by measuring samples in dilute conditions and crosschecking at multiple concentrations, thereby allowing a better estimate of the form factor alone to be obtained.

The commonly expected and often observed concentration effect in BioSAS is interparticle scattering, which manifests as concentration-dependent aggregation or repulsion (Fig. 21.3). These effects can be accounted for by continuing the dilution series of the measurements until no significant variation is observed or extrapolating the effect with concentration to infinite dilution (inferring the form factor at 0 mg/mL of protein in solution). However, any systematic over or underestimate of the concentration will lead to a



Name : run.xml
Type : STATIC
Date : Apr 16, 2015 10:47:31 AM

Comments: [BsxCube] Generated from BsxCube

Run	Frames (Average/Total)		Scattering	Concentration	Guinier			Gnom			Porod		Advanced
					Rg	Points	I(0)	Rg	Total	Dmax	Volume	MM Vol. est.	
# 42 # 41 # 40	hD	10/10 6/10 10/10		5.00 mg/ml	3.27 nm	51 - 76 (25)	21.71 ± 6.7e-2	3.43 nm	0.49	11.46 nm	57.57 nm³	28.8 - 38.4 kD	Data Reduction Abinitio Fit Superposition Rigid Body
# 44 # 43 # 42	hD	10/10 5/10 10/10		10.00 mg/ml	3.80 nm	27 - 50 (23)	23.81 ± 3.7e-2	3.81 nm	0.42	13.32 nm	67.56 nm³	33.8 - 45.0 kD	Data Reduction Abinitio Fit Superposition Rigid Body
# 47 # 46 # 45	hDu	10/10 10/10 10/10		1.00 mg/ml	2.41 nm	43 - 100 (57)	21.43 ± 7.9e-2	2.48 nm	0.47	8.07 nm	9.67 nm³	4.8 - 6.4 kD	Data Reduction Abinitio Fit Superposition Rigid Body
# 49 # 48 # 47	hDu	10/10 10/10 10/10		2.50 mg/ml	2.50 nm	38 - 95 (57)	19.36 ± 3.6e-2	2.56 nm	0.56	8.76 nm	32.20 nm³	16.1 - 21.5 kD	Data Reduction Abinitio Fit Superposition Rigid Body
# 51 # 50 # 49	hDu	10/10 8/10 10/10		5.40 mg/ml	2.56 nm	31 - 82 (51)	19.75 ± 2.3e-2	2.57 nm	0.53	8.58 nm	33.17 nm³	16.6 - 22.1 kD	Data Reduction Abinitio Fit Superposition

Fig. 21.4 Example summary from ISPyB, showing the automatically calculated invariants (Rg, I_0, D_{max}, Volume and MM estimates). Highlighted in *red* is the number of frames indicating the possibility of radiation damage present in these measurements and thus the need for manual checking of the curves, accessed via the Data Reduction button on the *right*. In the case a priori information is available in the form of known structures, comparisons direct to the data (including rigid body minimization in the case only separate domains are known) and overlayed with the ab-initio model can be viewed via the buttons to the *right* (*green* when available)

corresponding over or under-correction in the extrapolation.

Unfortunately, other factors besides interparticle scattering show concentration-dependent variations. In the case of weakly bound complexes, dilution may result in disassociation. Therefore, as concentration is reduced, so is the average size of the scatterers. To distinguish between these two possibilities careful checking of the MM estimates from both I_0 and Volume estimates are essential. In the case of interparticle scattering, as concentration is reduced the MM estimate will approach the expected MM, whereas in the case of dissociation it will appear below the expected MM for the complex.

A mixture might not only be the result of dissociation but also of an excess of one subunit from which the complex could not be purified. Like dissociation, as the subunit in excess is likely to be smaller than the complex (unless there are significant conformational changes on binding; see Sect. 21.5.5), then the MM estimates will be also underestimated for the mixture.

21.5.2 Considerations for Online Size Exclusion

Although the aim of performing online size exclusion chromatography (SEC) on mixtures is to separate the different species before collecting SAS data, sometimes the peaks will elute too close to one another. This can lead to overlapping peaks and in these regions the data measured will represent the mixed scattering from the overlapping species, with the proportions contributed to the total observed scattering by each species changing with time. It may still be possible to find regions where only one species predominates (checked by the presence of stable

Rg values), which can be merged together to give the scattering for that species. However, if no stable Rg can be found, direct merging of the data is not valid, as the underlying hypothesis of homogeneity and purity does not hold for a mixture of species. In the latter case, the sample needs to be remeasured using a better resolving column or, alternatively, deconvolution can be attempted to recover the scattering from the individual species [28].

An assumption in the data reduction of SEC data is that the scattering from the eluent preceding the elution of the peak is the background to be used for subtraction. This is in general true provided the column was fully equilibrated. Unfortunately UV is not an accurate measure especially when glycerol is present, as the signal is very small, and even if the UV signal is stable, variation could still be monitored by SAS. Therefore prior to injection of the sample, SAS baseline checks are recommended to ensure stability.

Another effect that can cause divergence from the background scattering is contamination of the sample holder. Some samples can stick to the surface of the measurement cell and result in additional scattering in subsequent frames. This effect is most readily observed as residual scattering (higher baseline) after the peaks. To attempt to correct for this, it is possible to apply a scaled subtraction using the assumption that the contamination is proportional to the concentration of the sample in the peak [17]. However, the validity of that assumption cannot be confirmed and, if the effect is strong, artifacts may still persist that should be taken into account in the interpretation of the data.

21.5.3 Flexibility

Some multi-subunit complexes may have an inherently high degree of conformational flexibility. An important consequence of the resulting structural heterogeneity is that the movement of the subunits in relation to each other will not be synchronized across all particles in the X-ray beam. Moreover, it can be assumed that all possible relative positions and orientations will be sampled in the scattering data under the assumption of spherical averaging (all possible orientations are present). This gives rise to an increase in the average size of the scatterers and, moreover, to variation in the particle sizes. These effects cause a deviation from the linear expectation of Guinier's law [13] and, as such, are practically indistinguishable in the 1D data from a small amount of aggregation. However, it is unlikely to have a dependence on concentration in the dilute concentrations for SAS experiments, though crowding effects can be observed in some cases at high concentrations (>10 mg/mL).

21.5.4 Verifying Stoichiometry

It is important to know with confidence the expected stoichiometry of the complex under study since reliably determining whether there are effects of dissociation or mixtures depends on comparisons of the MM estimate with the expected value. Therefore, isothermal titration calorimetry (ITC) experiments are highly recommended as a complementary method for the verification of the binding of any multi-subunit complex.

21.5.5 Partial Constructs

Additional information regarding the complex can be obtained from partial constructs. Even the additional information of the size of each part can be very informative to establish the stoichiometry and or presence of mixtures. If the volume of each construct is known, crosschecks can be applied to the complex, as the volume of the complex if fully bound will be an addition of the volumes of the bound subunits. If there is no binding or it is weak then the volume will be an average (in the case of mixture), or will lie between the average and the sum for a weakly bound complex.

It is therefore highly recommended when working with multi-subunit complexes to measure scattering data from all subunits and any possible partial constructs which can be obtained as well, in order to be able to better characterize, crosscheck and validate the results as well as to provide as much supporting evidence to increase confidence in the conclusions.

For example in order to determine the structure of the trimeric complex ABC (Fig. 21.5), six individual samples could be measured: three individual domains (Fig. 21.5a), two subunits (Fig. 21.5b, c), and the full complex (Fig. 21.5d). Note: in this example, domains A and C are not connected and therefore cannot be measured as a subunit. Thus it is important to know how the complex is formed to be able to measure as many individual subunits as possible. Additional functional studies can be planned measuring the complex ABC in different buffer conditions (Fig. 21.5d), containing ligands/additives required for activity and/or non-hydrolysable substrate analogs to isolate the various stages of the reaction. In this case, it is also interesting to measure the subunits and individual domains in all conditions. However, interpretation could be complicated as the individual domains alone may not be affected and only when two or more subunits are bound will there be any observable difference.

Partial constructs also provide additional scattering curves, which can be used simultaneously with the full complex data in *ab initio* or rigid-body modeling (see Sect. 21.6). However, all software programs intrinsically assume that the shape in the partial construct is the same in the complex – *i.e.*, that no conformational changes occur upon binding. This may in fact be the case but it should be verified experimentally, as it is possible that there are subtle or even dramatic [29] changes in the conformation upon binding of any or all subunits. It is possible that the differences are not necessarily a change in conformation but a difference in the intrinsic flexibility of the partial and fully bound constructs. Whatever the cause, any significant difference in the behavior (flexibility) or shape in different states (conformational change on binding) will have significant effects on the resulting models.

Understanding the nature of the samples and thus the validity of any analytic approach is essential. Furthermore, if differences are subtle, careful analysis is required to ensure that the variations are not artifacts and moreover that the differences are in fact statistically significant. Tools allowing statistical comparison of SAXS data are becoming available [30–32], which can provide quantitative and objective (superposition-independent) perspectives on solution state conformations.

Careful checks of the Rg, I_0 and volume should be undertaken to verify that the intrinsic assumption that there is no change in conformation during binding is valid, otherwise any model created will be uninformative. The volume and I_0 should always increase as the sum of the subunits if there is binding irrespective of changes in conformation. However, Rg is dependent on the shape and so can vary. Reduction in Rg while volume and I_0 increase is a very strong indication of compaction of the overall shape and significant conformational change on binding. If variations are found during this analysis, SANS using contrast variation with selectively labeled constructs can be used to obtain the shape of the individual subunits in the bound complex by matching out all other domains.

21.6 Modeling and Interpretation

The intrinsic ambiguity of *ab initio* modeling [33, 34] combined with its low resolution and the inherent issue that the modeling can be run on any curve even if it is not valid to do so, mean that all the crosschecks listed in the previous section must be performed and presented along with the modeling results for them to be accepted [35, 36]. Additionally, the presence of any partial constructs can be modeled independently as well as simultaneous modeling and then compared. Any significant variation between the results (displaced volume as well as overall shape) is an indication that the assumption of no conformational change on binding (as discussed above) is not valid.

21.6.1 Intrinsic Assumptions

Ab initio modeling requires intrinsically that the data used represents a monodisperse system.

Therefore, as working with multi-subunit complexes increases the likelihood of dissociation/mixtures (as discussed above), it is important to check the validity of the assumption of monodispersity and present confirmatory evidence (multiple SEC purifications, SEC MALS, DLS, AUC) to support the conclusions and validity of any models produced.

Another assumption is that of uniform contrast, posited by the majority of *ab initio* modeling programs, which use only two phases, one for the particle and a second one for the solvent, to represent the total scattering. The assumption of uniform contrast holds for conventional protein-only or nucleic acid-only samples, but breaks down for, *e.g.*, protein-DNA/RNA complexes, membrane proteins with bound lipid/detergent, or in neutron contrast variation. In the latter cases, the application of most *ab initio* modeling programs will yield biased results due to the assumption of uniform scattering contrast. Thus, the use of programs that allow multiphase *ab initio* modeling [34], is recommended for multi-subunit complexes.

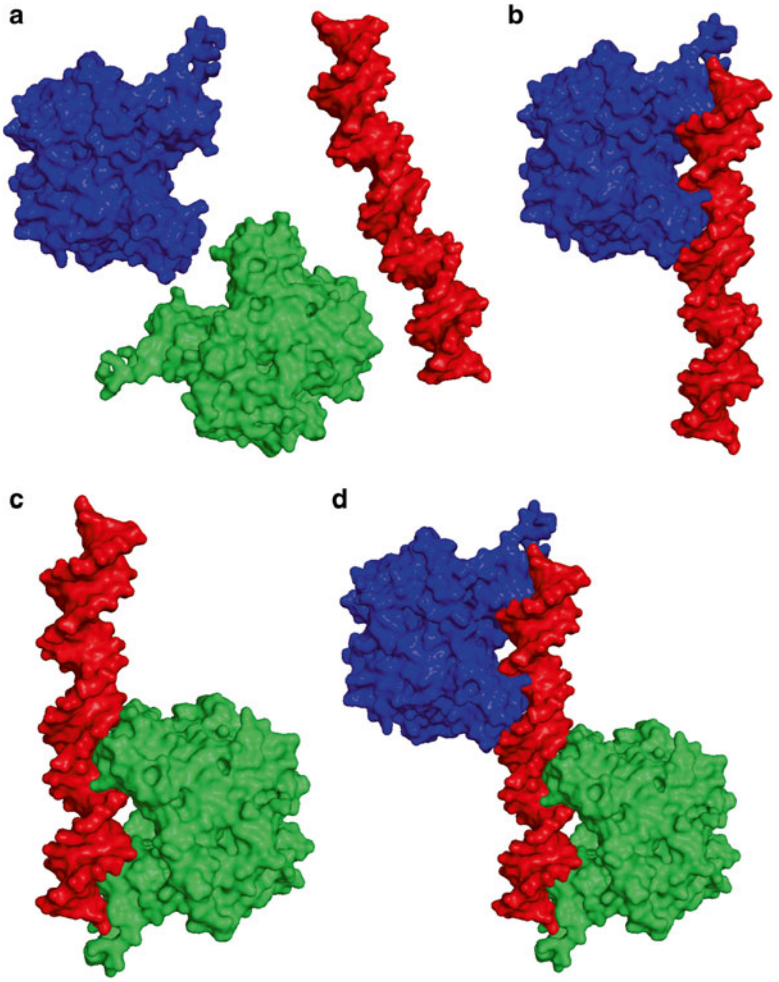


Fig. 21.5 Collecting SAXS data from full complex and partial constructs. Monomers A, B and C (**a**), the dimeric subcomplexes AB (**b**) and BC (**c**), and the full trimeric complex ABC (**d**). Samples can be measured in different conditions or contrasts (**e**)

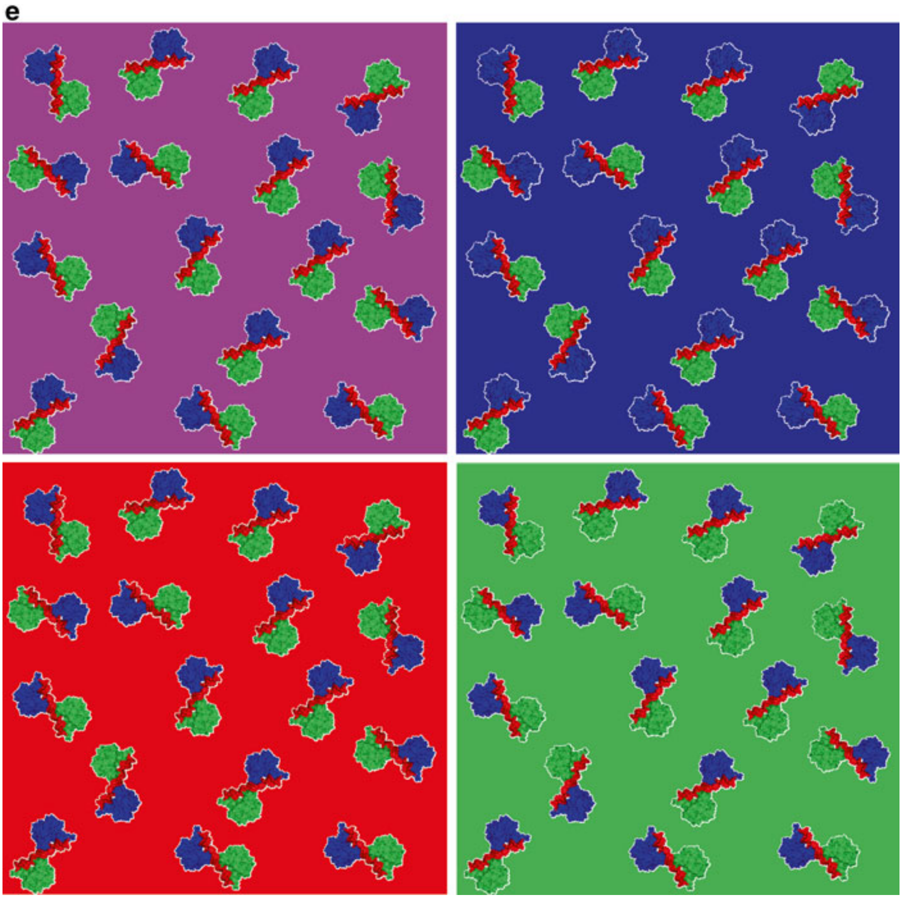


Fig. 21.5 (continued)

21.6.2 Inclusion of Information from Complementary Techniques

When studying multi-subunit complexes it is not uncommon for the structures of individual parts of the complex to have been determined by X-ray crystallography while the quaternary structure remains elusive because of flexibility, degradation/dissociation. It may also happen that complex binding is weak and therefore obtaining high-quality crystals becomes the limiting factor.

Likewise, NMR can be very useful to elucidate domain interactions inside subunits of a complex, since smaller domains may have been solved using NMR while larger complexes, due to intrinsic size constraints of the technique, are beyond the accessible range of NMR.

In the case of a high-resolution structure of the entire multi domain complex (or a homolog) is available; the theoretical scattering can be directly calculated and compared to the experimental data [37–40]. This comparison aids functional interpretation of the structure under different experimental conditions as well as helping to prevent crystallization artifacts effecting the conclusions. However, as the high resolution structures are individual snapshots, none of the obtained structures may represent the state in solution, in which case the domain positions from the structures can be altered and the modified model compared with the data. This process of altering the models and comparing to data is automated in a number of algorithms searching displacements along, and rotations about all three axes to find an optimal fit to the scattering data [41]. Furthermore this same process can be employed using domains solved separately with missing portions replaced by dummy residues [42]. However, in the case of flexibility, ensemble approaches [43, 44] may be required as a single model cannot accurately represent the scattering data. Additionally, the ensemble approaches can be used to assess the amount of flexibility present and thus provide useful crosschecks that flexibility is not significant and analysis using the standard *ab initio* and rigid-body modeling programs can proceed with confidence.

Although powerful tools for data analysis the rigid body approaches do not take into account any effects such as surface charge or hydrophobicity during the minimization, (although if such information is known it can be included as additional constraints). Molecular dynamic approaches in combination with BioSAS therefore provides a method to obtain models which are biologically relevant and compatible with the solution scattering data.

NMR can provide additional information (beyond structures to use in rigid body modeling) such as the relative orientations of domains which is difficult if not impossible (in the case of spherical domains) to obtain by solution scattering. Conversely, NMR is less sensitive to inter-domain distances which can be obtained relatively simply with BioSAS, a fact that is now being exploited using joint minimization [45, 46].

Electron microscopy (EM) can also provide very useful corroboration for multi-subunit complexes due to their size, although the individual domains may not be visualized alone. Comparison of the SAS models with the EM images or 3D reconstructions as well as crosschecks for size ensures compatibility and validity of biological conclusions [47, 48]. However, due to the intrinsic ambiguity of SAS reconstructions calculation of the theoretical scattering of the EM density map and comparison with the scattering data [42] may be preferable.

It is highly encouraged to use wherever possible complementary techniques to reduce ambiguity, increase confidence, and extend/strengthen the conclusions that can be made (see Table 21.3). However, care must be taken because any constraint used will bias the results and therefore the conclusions. Therefore, crosschecks and validation of all constrains/assumptions used must be made.

21.6.3 Data Analysis as in Iterative Process

Data analysis is not always straightforward; unfortunately, there will be times when you find that the initial assumptions are not valid. In a worst-case scenario, one may have to repeat an experiment. However, with the use of online data processing the necessity of repeat visits to synchrotrons is minimized as any required additional experiments or crosschecks are highlighted to alert the user immediately.

A major priority for data analysis is to validate the intrinsic assumptions applied by any modeling to ensure the result will be valid as well as any assumptions one makes regarding modeling constraints. Often the simplest way to check can be to run multiple modeling scenarios with different constraints and compare. Unfortunately, as there are many assumptions one may proceed with a modeling strategy only to discover it was not the best method and then one must start again.

Even the fundamental assumption of monodispersity must be verified and it may only be discovered that the sample is an equilibrium mixture following data acquisition. However, as mixtures can be analyzed with the use of complementary information [49], there may still be valuable biologically relevant results to be obtained, which will be the case if the mixture is the biologically relevant state.

Another common example of an assumption that has to be validated after data processing is symmetry. While the presence of true symmetry in a complex can be exploited to improve the

Table 21.3 Complementary techniques and information which can be used in modeling and analysis

Technique	Benefits	Issues	Recommendations
X-ray crystallography	High resolution structure of individual subunits/partial constructs	Crystal packing effects	Collect BioSAXS data from samples corresponding to all known constructs and compare to theoretical scattering from structures
		Conformation in solution may be different	Check for possibility of presence of mixtures
		Conformational changes on binding to other subunits	Careful crosschecking of invariants from subunits compared to partial constructs and whole complex
NMR	High resolution structure of individual subunits/partial constructs	Multiple conformations/ positions of side chains	Can give rise to errors in modeling if not accounted for
	Relative orientations of subunits	Limited size	Joint minimization using orientation constraints for NMR and distance from SAXS
EM	Complementary low resolution information on partial constructs/ quaternary structure	Non physiological conditions, sample preparation could induce artefacts	Thorough crosschecks of all available EM data. Are results uniform and reproducible?
	Direct comparison of EM reconstructions with BioSAXS. Calculation of theoretical scattering from EM volume compare to 1D BioSAXS data. Or real space overlays with *ab initio* models	Time consuming processing to obtain 3D structure	*Ab initio* models could be biased, therefore comparison of theoretical scattering to 1D curve is recommended
	BioSAXS data can be used to aid EM reconstruction	Possibility of bias in reconstruction	However, as the theoretical scattering curve is strongly affected by choice of threshold of EM volume a range of thresholds should be checked to assess the effect

quality of shape reconstruction by symmetry averaging (*e.g.*, P2, P3, P4), in certain cases only the core of a complex may obey the expected symmetry while other, more flexible subunits might break the symmetric arrangement. To test this, one should conduct modeling first in P1 (*i.e.*, assuming no symmetry) as well as with the appropriate symmetry. If the χ^2 values of the resulting models are not significantly different then the symmetry might be valid for the complex as well. However, if there is a significant deterioration in the quality of the fit with increasing symmetry then this is strong evidence that the symmetry does not hold for the full complex, and modeling strategies should be revised accordingly.

Moreover, it is necessary to compare the results of modeling using different strategies to ensure consistency as artifacts in the data manifest in different ways. For instance aggregation in *ab initio* modeling results in a larger model often with false extensions. However, for a rigid body model as the fit is dominated by the Rg, aggregation results in the distance between domains being increased. Thus rigid body models from data which suffers from the presence of aggregation will tend to be more extended. Overlays of *ab initio* and rigid body models from the same data should ideally superimpose, and therefore any mismatch is a strong indication of either a false assumption in the modeling constraints or artifacts in the data.

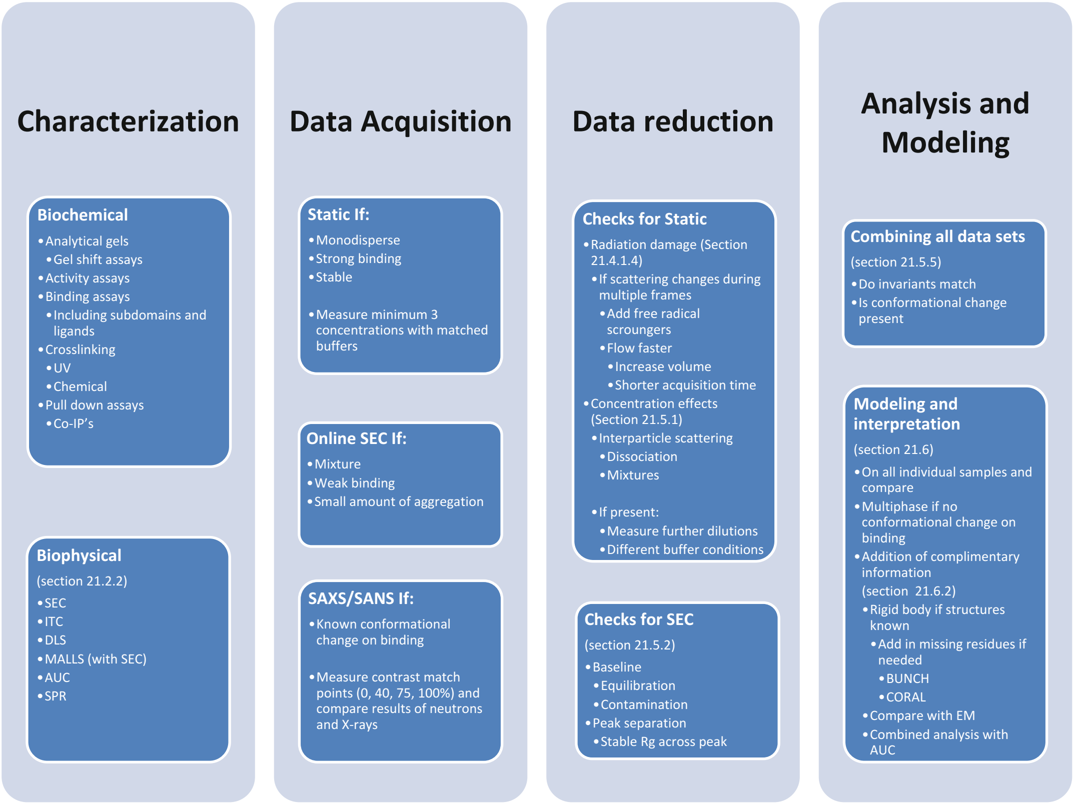
One should not be afraid to try any and all possibilities for modeling including the addition of constraints, in order to understand not only which is best but also the extent of the conclusions that can be made. As such, data analysis should be viewed as part of an exploratory, interactive process designed to test hypotheses, validate conclusions, and improve the models that are produced.

21.7 Concluding Remarks

BioSAXS is a very powerful tool for the study of multi-subunit complexes as it not only allows to bring together complementary information across different scales, from biochemical data to high-resolution atomic structures, but it also does so under physiologically relevant conditions in solution. This means that not only the quaternary structure in solution can be obtained but that functional studies of the subunits and full complex are also possible.

Although BioSAXS is a relatively simple experiment, complexity arises from the samples under investigation, which for multi-subunit complexes depends on the number of subunits and their behavior.

As highlighted in this chapter, care must be taken when performing BioSAXS experiments, careful planning and execution of data acquisition with thorough data analysis and crosschecks (summarized in Fig. 21.6) are essential to provide insight and the elucidation of complex biological processes of great interest to the wider scientific population and general public.



Characterization

Biochemical
- Analytical gels
  - Gel shift assays
- Activity assays
- Binding assays
  - Including subdomains and ligands
- Crosslinking
  - UV
  - Chemical
- Pull down assays
  - Co-IP's

Biophysical
(section 21.2.2)
- SEC
- ITC
- DLS
- MALLS (with SEC)
- AUC
- SPR

Data Acquisition

Static If:
- Monodisperse
- Strong binding
- Stable
- Measure minimum 3 concentrations with matched buffers

Online SEC If:
- Mixture
- Weak binding
- Small amount of aggregation

SAXS/SANS If:
- Known conformational change on binding
- Measure contrast match points (0, 40, 75, 100%) and compare results of neutrons and X-rays

Data reduction

Checks for Static
- Radiation damage (Section 21.4.1.4)
  - If scattering changes during multiple frames
    - Add free radical scroungers
    - Flow faster
      - Increase volume
      - Shorter acquisition time
- Concentration effects (Section 21.5.1)
  - Interparticle scattering
    - Dissociation
    - Mixtures
  - If present:
    - Measure further dilutions
    - Different buffer conditions

Checks for SEC
(section 21.5.2)
- Baseline
  - Equilibration
  - Contamination
- Peak separation
  - Stable Rg across peak

Analysis and Modeling

Combining all data sets
(section 21.5.5)
- Do invariants match
- Is conformational change present

Modeling and interpretation
(section 21.6)
- On all individual samples and compare
- Multiphase if no conformational change on binding
- Addition of complimentary information
  (section 21.6.2)
  - Rigid body if structures known
    - Add in missing residues if needed
      - BUNCH
      - CORAL
  - Compare with EM
  - Combined analysis with AUC

Fig. 21.6 Overview of SAS experimental workflow

Acknowledgments I would like to thank the EMBL and ESRF for supporting BioSAXS with both funds and manpower. I would also like to thank all colleagues at the EMBL and the ESRF and the PSB in general for all the discussions and complimentary work which has not only helped in writing this but in my research in general. Special thanks to A. De Maria Antolinos for help with Fig. 21.4, to the editor for help and feedback to this chapter and to my family for their continual support in my work in general and specifically during the writing process.

References

1. Yu H (1999) Extending the size limit of protein nuclear magnetic resonance. Proc Natl Acad Sci U S A 96:332–334
2. Skrisovska L, Schubert M, Allain FH-T (2010) Recent advances in segmental isotope labeling of proteins: NMR applications to large proteins and glycoproteins. J Biomol NMR 46(1):51–65
3. Hong Zhou Z (2011) Recent advances in electron cryomicroscopy, part B. Chapter 1: atomic resolution cryo electron microscopy of macromolecular complexes. Adv Protein Chem Struct Biol 82:1–35
4. Kuhlbrandt W (2014) Microscopy: cryo-EM enters a new era. eLife 3, e03678
5. Bai X-C, McMullan G, Scheres SHW (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40(1):49–57
6. Lu P, Xiao-chen Bai X-C, Ma D, Xie T, Yan C, Sun L, Yang G, Zhao Y, Zhou R, Scheres SHW, Shi Y (2014) Three-dimensional structure of human Î³-secretase. Nature 512:166–170
7. McPherson A (2004) Introduction to protein crystallization. Macromolecular crystallization. Methods 34(3):254–265
8. Yu L, Reutzel-Edens SM, Mitchell CA (2000) Crystallization and polymorphism of conformationally flexible molecules: problems, patterns, and strategies. Org Proc Res Dev 4(5):396–402
9. Janin J, Rodier F (1995) Protein-protein interaction at crystal contacts. Proteins Struct Funct Genet 23:580–587
10. Hinsen K (2008) Structural flexibility in proteins: impact of the crystal environment. Bioinformatics 24(4):521–528
11. Rupp B (2010) Biomolecular crystallography: principles, practice and application to structural biology. Garland Science, New York, p 83
12. Zerrad L, Merli A, Schroder GF, Varga A, Graczer E, Pernot P, Round A, Vas M, Bowler MW (2011) A spring-loaded release mechanism regulates domain movement and catalysis in phosphoglycerate kinase. J Biol Chem 286:14040–14048
13. Guinier A (1938) The diffusion of X-rays under the extremely weak angles applied to the study of fine particles and colloidal suspension. Comptes Rendus Hebdomadaires Des Seances De L Acad Des Sci 206:1374–1376
14. Porod G (1982) In: Glatter, Kratky (eds) Small angle X-ray scattering. Academic, London
15. Glatter O (1977) Anew method for the evaluation of small angle scattering data. J Appl Crystallogr 10:415–421
16. Svergun DI (1992) Determination of the regularisation parameter in indirect-transform methods using perceptual criteria. J Appl Crystallogr 25:294–503
17. Round AR, Franke D, Moritz S, Huchler R, Fritsche M, Malthan D, Klaering R, Svergun DI, Roessle M (2008) J Appl Crystallogr 41:913–917
18. Blanchet CE, Zozulya AV, Kikhney AG, Franke D, Konarev PV, Shang W, Klaering R, Robrahn B, Hermes C, Cipriani F, Svergun DI, Roessle M (2012) Instrumental setup for high-throughput small- and wide-angle solution scattering at the X33 beamline of EMBL Hamburg. J Appl Crystallogr 45:489–495
19. David G, Pérez J (2009) J Appl Crystallogr 42:892–900
20. Classen S, Hura GL, Holton JM, Rambo RP, Rodic I, McGuire PJ, Dyer K, Hammel M, Meigs G, Frankel KA, Tainer JA (2013) Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the advanced light source. Appl Crystallogr 46(Pt 1):1–13
21. Pernot P, Round A, Barrett R, De Maria Antolinos A, Gobbo A, Gordon E, Huet J, Kieffer J, Lentini M, Mattenet M, Morawe C, Mueller-Dieckmann C, Ohlsson S, Schmid W, Surr J, Theveneau P, Zerrad L, McSweeney S (2013) J Synchrotron Radiat 20:660–664
22. Pernot P, Theveneau P, Giraud T, Nogueira-Fernandes R, Nurizzo D, Spruce D, Surr J, McSweeney, Round AS, Felisaz F, Foedinger L, Gobbo A, Huet J, Villard C, Cipriani F (2010) J Phys Conf Ser 247:012009
23. Round A, Brown E, Marcellin R, Kapp U, Westfall CS, Jez JM, Zubieta C (2013) Acta Cryst D69:2072–2080
24. Schoenborn B (1975) Neutron scattering for the analysis of biological structures. Brookhaven National Laboratory, Upton, pp 110–117
25. Garman EF (2010) Radiation damage in macromolecular crystallography: what is it and why should we care? Acta Crystallogr D Biol Crystallogr 66(Pt 4):339–351
26. Koch MHJ, Sayers Z, Sicre P, Svergun D (1995) Macromolecules 28:4904–4907
27. Franke D, Kikhney AG, Svergun DI (2012) Automated acquisition and analysis of small angle X-ray scattering data. Nucl Instrum Methods Phys Res, Sect A 689:52–59
28. Brookes E, Perez J, Cardinali B, Profumo A, Vachette P, Rocco M (2013) Fibrinogen species as resolved by HPLC-SAXS data processing within the UltraScan SOlution MOdeler (US SOMO) enhanced SAS module. J Appl Crystallogr 46:1823–1833

29. Koehler C, Round A, Simader H, Suck D, Svergun DI (2013) Quaternary structure of the yeast Arc1p-aminoacyl-tRNA synthetase complex in solution and its compaction upon binding of tRNAs. Nucleic Acids Res 41(1):667–676
30. Grant TD, Luft JR, Carter LG, Matsui T, Weiss TM, Martel A, Snell EH (2015) The accurate assessment of small-angle X-ray scattering data. Acta Crystallogr D Biol Crystallogr 71(Pt 1):45–56
31. Hura GL, Budworth H, Dyer KN, Rambo RP, Hammel M, McMurray CT, Tainer JA (2013) Comprehensive objective maps of macromolecular conformations by quantitative SAXS analysis. Nat Methods 10(6):453–454
32. Franke D, Jeffries CM, Svergun DI (2015) Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. Nat Methods 12:419–422
33. Franke D, Svergun DI (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Crystallogr 42:342–346
34. Svergun DI (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys J 76:2879–2886
35. Jacques DA, Trewhella J (2010) Small-angle scattering for structural biology: expanding the frontier while avoiding the pitfalls. Protein Sci 19(4):642–657
36. Jacques DA, Guss JM, Svergun DI, Trewhella J (2012) Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. Acta Crystallogr D Biol Crystallogr 68(Pt 6):620–626
37. Svergun D, Barberato C, Koch MHJ (1995) CRYSOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 28:768–773
38. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2013) Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophys J 105(4):962–974
39. Schneidman-Duhovny D, Hammel M, Sali A (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res 38(Suppl):W540–W544
40. Knight CJ, Hub JS (n.d.) WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. Nucl Acids Res. doi: 10.1093/nar/gkv309
41. Petoukhov MV, Svergun DI (2005) Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophys J 89:1237–1250
42. Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HDT, Konarev PV, Svergun DI (2012) New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Cryst 45:342–350
43. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc 129(17):5656–5664
44. Pelikan M, Hura GL, Hammel M (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys 28:174–189
45. Gabel F, Simon B, Nilges M, Petoukhov M, Svergun D, Sattler M (2008) A structure refinement protocol combining NMR residual dipolar couplings and small angle scattering restraints. J Biomol NMR 41(4):199–208
46. Lapinaite A, Simon B, Skjaerven L, Rakwalska-Bange M, Gabel F, Carlomagno T (2013) The structure of the box C/D enzyme reveals regulation of RNA methylation. Nature 502:519–523
47. Alcorlo M, Martínez-Barricarte R, Fernández FJ, Rodríguez-Gallego C, Round A, Vega MC, Harris CL, Rodríguez de Cordoba S, Llorca O (2011) Unique structure of iC3b resolved at a resolution of 24 Å by 3D-electron microscopy. Proc Natl Acad Sci U S A 108(32):13236–13240
48. Sulák O, Cioci G, Lameignère E, Balloy V, Round A, Gutsche I, Malinovská L, Chignard M, Kosma P, Aubert DF, Marolda CL, Valvano MA, Wimmerová M, Imberty A (2011) Burkholderia cenocepacia BC2L-C is a super lectin with dual specificity and proinflammatory activity. PLoS Pathog 7(9), e1002238
49. Konarev PV, Volkov VV, Sokolova AV, Koch MHJ, Svergun DI (2003) PRIMUS – a windows-PC based system for small-angle scattering data analysis. J Appl Crystallogr 36:1277–1282

22 Application of Nuclear Magnetic Resonance and Hybrid Methods to Structure Determination of Complex Systems

Filippo Prischi and Annalisa Pastore

Abstract

The current main challenge of Structural Biology is to undertake the structure determination of increasingly complex systems in the attempt to better understand their biological function. As systems become more challenging, however, there is an increasing demand for the parallel use of more than one independent technique to allow pushing the frontiers of structure determination and, at the same time, obtaining independent structural validation. The combination of different Structural Biology methods has been named hybrid approaches. The aim of this review is to critically discuss the most recent examples and new developments that have allowed structure determination or experimentally-based modelling of various molecular complexes selecting them among those that combine the use of nuclear magnetic resonance and small angle scattering techniques. We provide a selective but focused account of some of the most exciting recent approaches and discuss their possible further developments.

Keywords

Multidomain proteins • NMR • Protein-protein interactions • Structural biology • SAXS

22.1 Introduction

Not so long ago researchers would be mainly experts in one field and carry out their research using a very specific expertise. Nowadays, the number of techniques at hand has enormously increased enhancing the possibility of independently validating results with more than one tool. This is particularly true in Structural Biology, a field that has gained enormous momentum in the

F. Prischi
School of Biological Sciences, University of Essex,
Wivenhoe Park, Colchester CO4 3SQ, UK

A. Pastore (✉)
Department of Clinical Neurosciences, King's
College London, Denmark Hill Campus, London, UK
e-mail: Annalisa.Pastore@crick.ac.uk

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*,
Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0_22

post-genomic era. The possibility of using more than one technique has also suggested new approaches, which may allow one to combine results and obtain a better and more complete picture thus moving further the frontiers of structure determination. In this review, we will focus on so-called 'hybrid' techniques developed for solution studies. We will first briefly overview what can by now be considered classical methods based on nuclear magnetic resonance (NMR) techniques to then overview hybrid methods combining NMR with small angle X-ray scattering (SAXS). This technique has proven well suited to provide information on the overall shape of a molecule and on the non-uniform distribution of the protein atomic density [1]. Because of their complementarity, NMR being a higher resolution technique but unable to deliver information on big assemblies, SAXS being low resolution but very effective in the reconstruction of the overall shapes, the two techniques are often exploited in combination, as detailed in the next pages.

There are currently three main applications of hybrid methods based on these two techniques: the definition of the relative orientation of multidomain proteins, structure refinement of proteins having sparse distance restraints and the reconstruction of large molecular complexes [2]. We will review some examples, discuss the limitations encountered and suggest new directions.

22.2 Why NMR Cannot Do Everything

The time of structural studies of single domain proteins has rapidly had its sunset. Most of the forefront current structural biology projects deal with multi-domain proteins and large molecular assemblies, which can be as big as or larger than the ribosome [3, 4]. These changed priorities have revolutionized our perspective of the tools needed for structure determination. A well-established way, which quickly approaches its 30th birthday, is to use the so-called cut-and-paste approach that relies on the minimalistic approach of cutting a protein/complex into isolated domains/components and solving their structures and their interactions in isolation. Only after these are studied, we will want to find ways to determine the relative orientation of the individual parts. Both the two main techniques traditionally used for structure determination, *i.e.*, X-ray crystallography and nuclear magnetic resonance (NMR) in solution, have strongly benefitted from this concept although for different reasons.

The cut-and-paste approach is good in crystallography because it allows cutting away, at least in a first instance, flexible regions, which could be difficult to crystallize. It is also helpful in NMR studies: structure determination with classic NMR methods solely based on nuclear Overhauser effects (NOEs) can be very challenging with large proteins, because, due to slower rotational diffusion, the line widths increase, resulting in a decrease of the signal-to-noise ratio and an increase of resonance overlap up to the disappearance of the signals [5]. The exact limit is not simply a function of the molecular weight: proteins with similar molecular weights can be observed or not according to whether they are intrinsically unfolded or rigidly globular. The limit can anyway be extended by uniformly or selectively deuterating most of the molecule's protons, which effectively 'dilutes' out the spin concentration and increases the T_2 transversal relaxation with consequent decrease of the line widths [6]. Selective labeling, perdeteuration and the use of TROSY pulse sequences in high-field NMR spectrometers (900 MHz or higher fields) have dramatically increased the signal-to-noise ratio [7].

Besides the molecular weight, another limitation of NMR is encountered when wanting to determine the relative orientation of multi-domain proteins. The task is particularly problematic when systems do not have a rigid and extended interface. This is because NOEs are intrinsically short-range observables [5] (it could of course be argued that alternative techniques are not necessarily much better: X-ray crystallography can well obtain long-range information but the suspect can be that the result might be biased by the very process of crystallization). Residual dipolar couplings (RDCs) were introduced to resolve the problem. Internuclear magnetic dipole couplings contain a great deal of structural information, but they average to zero in isotropic solution as a result of rotational diffusion [8]. Tjandra and Bax developed a method in which alignment of proteins with the magnetic field can be achieved through the use of weak liquid crystalline (LC) media [9]. This induces an anisotropic distribution of orientations that allows accurately measurement of a wide array of RDCs [9] which provide a powerful source of orientational restrains by defining the angles between an inter-nuclear vector and the axis of an alignment frame. The alignment frame works as an external reference and is fixed to the molecular frame of the molecule. An advantage of this formalism is that, since RDCs are relative to the molecular frame, they are independent from the tumbling of the molecule, hence they pick up motions faster and slower than rotational tumbling correlation time of the molecule. This makes RDCs powerful tools to monitor protein dynamics.

The major disadvantage when using orientational restrains in the construction of oligomer models is that for a single set of RDCs and a structural model exists four combinations of relative orientation of the subunits in the complex [10]. This uncertainty can be resolved by collecting a second (or more) data set of RDCs from a different alignment medium [10]. Not always this is possible though as not all media are suitable for all proteins. Another way to get around these limitations is to combine RDCs with other NMR observable and builds models from these hybrid restraints [11–13]. These include not only NOEs, but also paramagnetic relaxation enhancement (PRE) and chemical shift perturbation (CSP) data. SAXS has been used as an alternative or in combination with these methods.

22.3 SAXS in Defining Multi-domain Relative Orientations

As illustrated by Mertens and Svergun, the analysis of flexible systems by SAXS has received a great boost by Bernado and collaborators [14]. A recent study [15] focused on the difficulties associated with the interpretation of SAXS curves for highly flexible proteins. These proteins are at times identified as rigid from dynamically averaged SAXS profiles unless several indicators are monitored. The best approach resides in a method called ensemble optimization method (EOM) [16] because it provides a reliable measure of the flexibility of the system under study. In the EOM approach it is necessary to generate first a large pool of random configurations and then select ensembles using a genetic algorithm [16]. When the Rg distribution of the models in the selected ensembles is as broad as that in the initial random pool, the protein is probably flexible, whereas a narrow Rg peak hints at a rigid system. The combination of SAXS and NMR spectroscopy provides averages of the entire ensembles of conformation [17–20]. However, it is very challenging to identify consistent ensembles, given the vast number of conformations that can potentially be adopted by flexible proteins. Several approaches have been developed.

Probably the first examples of using SAXS in combination with NMR information were published back in 1996–1997 in studies aiming at reconstructing modular proteins from the individual domain [21]. Sunnerhagen et al. [21] used this combination of techniques to study the relative orientation of Gla and EGF domains in the coagulation factor X. This is a serine protease containing three noncatalytic domains: an N-terminal gamma-carboxyglutamic acid (Gla) domain followed by two epidermal growth factor (EGF)-like domains. It was noticed that when linked to the Gla domain, the Ca^{2+} affinity of the isolated N-terminal EGF domain is increased tenfold suggesting a cross-talk between the two domains. Through a study of the NMR solution structure of the factor X Gla-EGF domain pair (with Ca^{2+} bound to the EGF domain), complemented by SAXS data on the Gla-EGF domain pair (with and without Ca^{2+}), the authors showed that Ca^{2+} binding to the EGF domain makes the Gla and EGF domains fold toward each other using the Ca^{2+} site as a hinge. Presumably, a similar mechanism may be responsible for alterations in the relative orientation of protein domains in many other extracellular proteins containing EGF domains with the consensus for Ca^{2+} binding. Finally, this study demonstrated the powerful combination of NMR and SAXS in the study of modular proteins, since it combines reliable evaluation of short- (NMR) and long-range (SAXS) interactions.

Our group used NMR/SAXS shortly after to build a model of how the multidomain protein titin is assembled [22, 23]. Titin is a giant muscular protein and a prototype of a modular protein containing *ca*. 300 copies of two all-β sequence motifs, the fibronectin type 3 and the immunoglobulin-like modules [23]. An important question was (and still to some extent is) whether titin modules interact with each other or are loosely connected without intermodule interactions. The question was addressed by assessing the extent of CSP between modules and measuring by SAXS the maximal distance of constructs of two- and four-modules. It was concluded that the linkers connecting the domains in the I-band are relatively rigid and dictate a total length of the multi-domain constructs shorter than the one expected for the sum of the individual domains.

Bertini et al. [19] developed an algorithm to determine the maximum occurrence (MO) of a given conformation, or the maximum percent of time a system spends in a given conformation. The program, publicly available using the grid computing infrastructure (https://www.wenmr.eu/), initially generates a pool of about 10^5 conformations using RANCH [20]. Theoretical NMR and SAXS data are generated for each conformation, using FANTASIAN [24] and CALCALL [19] to estimate respectively pseudo-contact shifts (PCS) and residual dipolar couplings (RDC) and using CRYSOL [25] to calculate SAXS intensities. At this point, a conformation (A) is selected and assigned a weight lower than 100 %. A group of other conformations (randomly selected from the initial pool) is added to this conformation with a weight that is adjusted to obtain the best fit between experimental and theoretical data. The program varies not only the weight of the different conformations, but also discards and substitutes other conformations from the pool, with the selection driven by a simulated annealing protocol. The procedure stops when the target function (TF) reaches a minimum value and the MO for A is determined (Fig. 22.1a, b). This methodology was applied to calmodulin (CaM), a classic model of a flexible two-domain protein [reviewed in 26]. The MO of the Ca^{2+} bound (PDB ID: 1CLN, 1CLL, 1PRW) [27–29] and of the closed peptide-bound forms (PDB ID: 1CDL, 1CDM, 1IQ5, 1NIW, 1YR5, 2BCX, 2XOG) was evaluated [30–35]. It was concluded that dumbbell-shape extended conformations as well as compact conformations have very low occupancy (MOs in the order of 5–15 %). More expanded (or, implicitly, more flexible) conformations have MO as high as 35 %, strongly suggesting that these conformations are most abundant in solution. These results mainly confirmed decades of previous studies that had already established the absence of extended conformations of calmodulin in solution [36–38].

A more sophisticated and interesting approach was implemented by Huang et al. [39] using a combination of PRE, RDC and SAXS data to study U2AF65. This protein is essential for spliceosome assembly [40] and is composed by three RNA Recognition Motif (RRM) domains connected by flexible linkers. PRE studies revealed a dynamic equilibrium between a predominant compact "closed" conformation and a less abundant "open" RNA-bound-like conformation [40]. This is a key feature in the recognition of a range of polypyrimidine (Py) tracts found in human pre-mRNA introns [40]. However, structures of the open and closed conformations do not completely fit SAXS data, indicating that the range of conformations sampled by U2AF65 in solution is much wider. To study these large-scale dynamic modes that are known to play key roles in a multitude of molecular recognition and signaling processes, the authors used the software ASTEROIDS [41, 42]. The software maps the

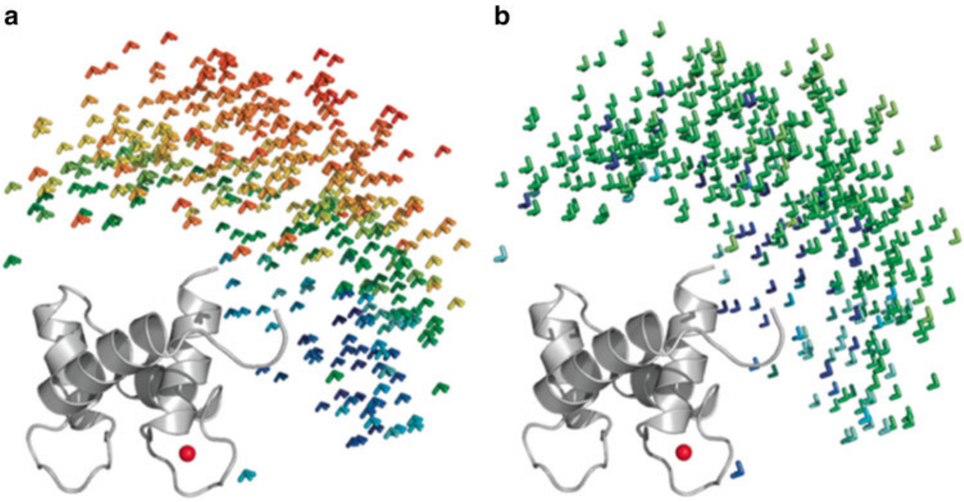


Fig. 22.1 Orientation tensors centered in the center-of-mass of the C-terminal domain of CaM are color-coded with respect to the MO of the corresponding conformation, from *blue* (<5 %) to *red* (>40 %). Two different orientations (panels **a**, **b**) of the tensors are chosen to show that MO depends on both the relative domain orientation and the position. A high- (**a**) and a low-MO orientation (**b**) are chosen (Reprinted with permission from Bertini et al. [19]. Copyright 2010 American Chemical Society)

conformational space adopted by the protein in an unbiased way using a sequence-dependent stochastic sampling algorithm. Experimental data are used to select from this pool ensembles of conformations. Instead of optimizing the weight of a conformer in the ensemble, the software uses the genetic algorithm to increase the number of copies of a specific conformer in the ensemble in order to obtain a better fitting of the experimental data.

This is conceptually different from the procedure adopted by Bertini et al. [19]. The difference in parameterization allows ASTEROIDS to perform a robust noise-based Monte Carlo error analysis, independently from the quality of the experimental data. The authors used this approach to map the conformational energy surface of the first two RRM domains of U2AF65, inputting RDC, PRE and SAXS data directly into ASTEROIDS. They concluded that the two domains are mainly in an extended conformation while the previously reported “closed” and “open” forms [40] are adopted only by one-quarter of conformers (mostly in the “closed” form). The authors rationalize their results by suggesting that, even if the structures of the “open” and “closed” forms differ appreciably, they lie within a continuous ensemble envelope and this could represent a possible “pathway” of available states that can flow between the two forms without major expensive energetic jumps.

22.4 Hidden Interdomain Information: Direct Structure Refinement

While reducing spin diffusion, perdeuteration also reduces the number of the resonances in the spectrum, including the majority of the resonances necessary for evaluating NOE effects between interdomain side-chains. In the attempt to obtaining experimental information that could compensate for this loss of information, Bax and co-workers [43] implemented SAXS data in NMR structure refinement. As in other SAXS applications, χ^2 statistics were used to evaluate back-calculated scattering curves during the molecular dynamic/energy minimization steps. The cycles of structure refinement were stopped when convergence, *i.e.*, the minimum of χ^2, was reached. To make this procedure “slimmer” from a time and computational point of view, the authors used a “globic approximation”, *i.e.*, they represented the protein structure by small fragments of 3–9 heavy atoms (previously applied for X-ray crystallography [44] and SAXS [45, 46] structure determination) and the SAXS curve by a limited number of data points. The SAXS data-fitting module was implemented into the CNS structure refinement package [47]. Since the first publication, several structures have been solved following this approach (PDB ID: 2A5M, 2JQX, 2K4C, 2KX9, 2XDF, 2L5H) [43, 48–51]. Bax and co-workers tested this approach on γS-Crystalline [43], malate synthase G (MSG) [48] and on tRNA^{Val} [49].

γS-Crystalline is a two-domain protein of 177 residues. The NMR structure of the protein is very similar to that of other homologs (each domain consists of two four-strand β-sheets arranged in Greek key motifs) solved by X-ray crystallography. However, the orientation of the two domains in the NMR structure differs significantly due to lack of interdomain NOE restrains. Refinement including SAXS data allows a better agreement with the structures of other orthologs. The two domains pack closer to each other with a better backbone rmsd as compared to the γB- and γD-crystalline [52, 53] crystal structure (respectively without and with SAXS data 1.96–1.31 Å and 1.07–0.87 Å). No major translations are detected and only 5.5° rotation of the C-terminal domain is present when N-terminal domain is used for structure alignment (Fig. 22.2).

Malate synthase G (MSG) [48] is a challenging example, since it is an 82 kDa protein and currently the largest single chain protein solved by solution NMR. MSG catalyzes the chemical reaction between acetyl-CoA and glyoxylate to form malate and CoA. The structure of the enzyme was solved both by X-ray crystallography (PDB ID: 1D8C, 1N8I) [54, 55] and NMR (PDB ID: 1Y8B) [56]. The basic fold of the enzyme is that of a β8/α8 (TIM) barrel with an N-terminal α-helical domain flanking one side and a C-terminal α-helical domain forming a plug which caps the active site and a α/β domain with unknown function. Similarly to the γS-crystalline structure, the inclusion of SAXS data improved the structure refinement and allowed an overall improvement of the backbone root mean square deviation (rmsd) compared to the crystal structure (PDB ID: 1D8C) from 4.92 to 1.39 Å. A translation of ~4°–5° for α/β domain and ~3°–4° for the C-terminal domain between the NMR-only and NMR/SAXS-refined structures was observed (Fig. 22.3). Globular domains like the β8/α8 TIM barrel benefit the most from the introduction of SAXS data to counterbalance the reduction of information consequent to perdeuteration.

Clore and co-workers [57] used a combination of NMR and SAXS data for the study of the full length HIV-1 capsid protein, a challenging system that had given a hard time to several structural biologists. The HIV-1 capsid is a key component in viral infection. It is composed of N- and C- terminal domains connected by a flexible linker (Fig. 22.4a) with the N-terminal domain forming hexameric and pentameric rings (Fig. 22.4b) and the C-terminal domain forming homodimers that connect adjacent N-terminal domain rings [58–62] (Fig. 22.4c). The main problem encountered was caused by the backbone resonances of the linker residues and of residues at the dimer interface of full length HIV-1 capsid protein, which are broad because of monomer/dimer exchange. Similarly to the other two methodologies described above, the authors proceeded with mapping the conformational space sampled by the N-terminal domain relative to the C-terminal domain using a RDC and SAXS/WAXS-driven simulated annealing [50, 63]. It was noticed that the relative orientation of the N- and C-terminal domains does not overlap between the monomeric and dimeric forms and hence the authors postulated that oligomerization acts as a modulator of orientation equilibrium. Interestingly, intra-subunit interactions were detected in the monomeric form. These interactions were driven by the accessible hydrophobic dimerization helix (residues 179–192) of the C-terminal domain that makes contact with residues of the N-terminal domain (Glu29 and Ala31) and the linker region (highlighted by a circle in Fig. 22.4c). On the other hand, the HIV-1 capsid protein dimer is characterized by a single orientation of the C-terminal domain that is in agreement with the previously solved NMR

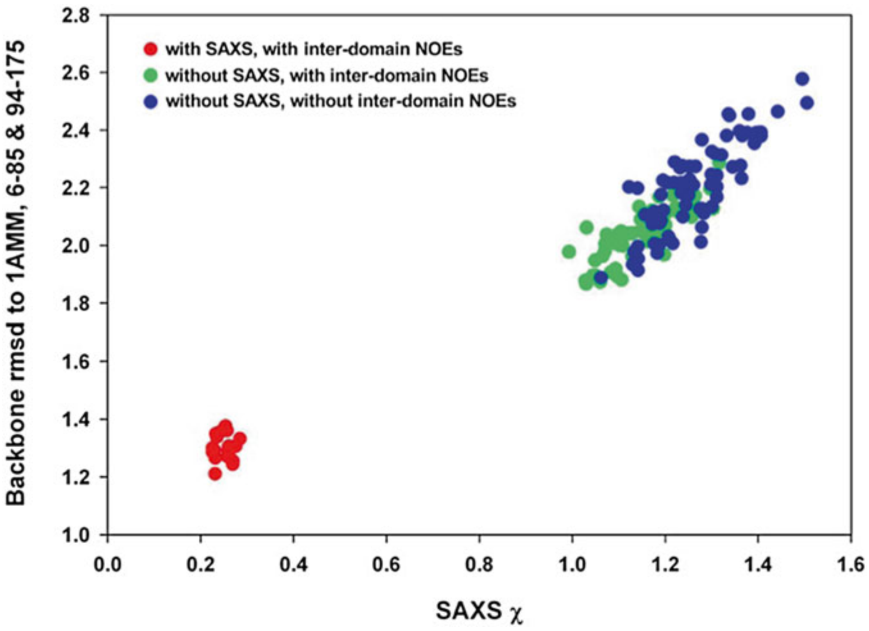


Fig. 22.2 Plot of the correlation between SAXS χ and backbone rmsd to 1AMM (residues 6–85 and 94–175). These results show that calculation of a family of structures with the SAXS data produces an improvement in the structural accuracy [43] (Reprinted with permission from Grishaev et al. [43] Copyright 2005 American Chemical Society)

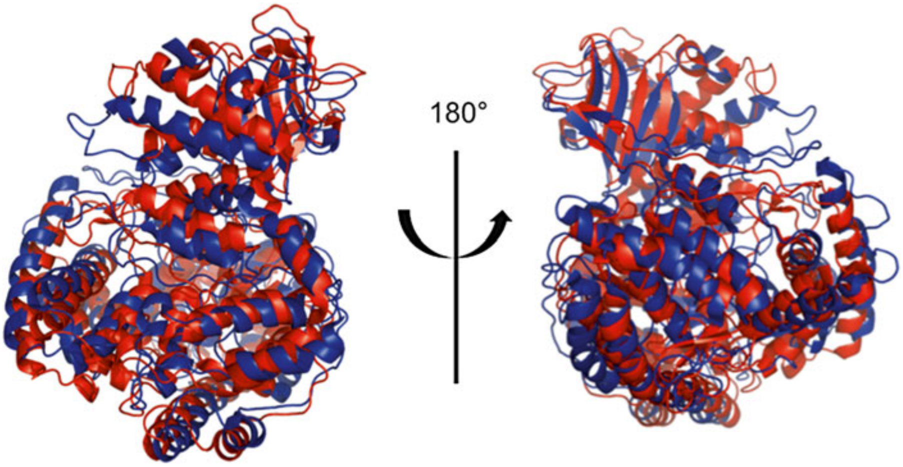


Fig. 22.3 Structural superposition of malate synthase G obtained by the joint fit of SAXS and NMR data (*red*, PDB ID: 2JQX) [48] and the NMR-only model (*blue*, PDB ID: 1Y8B) [56]

structure [64] and in contrasts with the crystal structure [65, 66]. Dimerization of the HIV-1 capsid protein prevents the formation of intra-subunit interactions. This information is important for HIV treatment. Hydrophobic capsid assembly inhibitors [67, 68] stabilize, through hydrophobic interactions, the interaction of the N-terminal domain with the C-terminal one shifting the monomer-dimer equilibrium in favor of the monomer preventing interaction between pentameric or hexameric assemblies and blocking capsid formation.

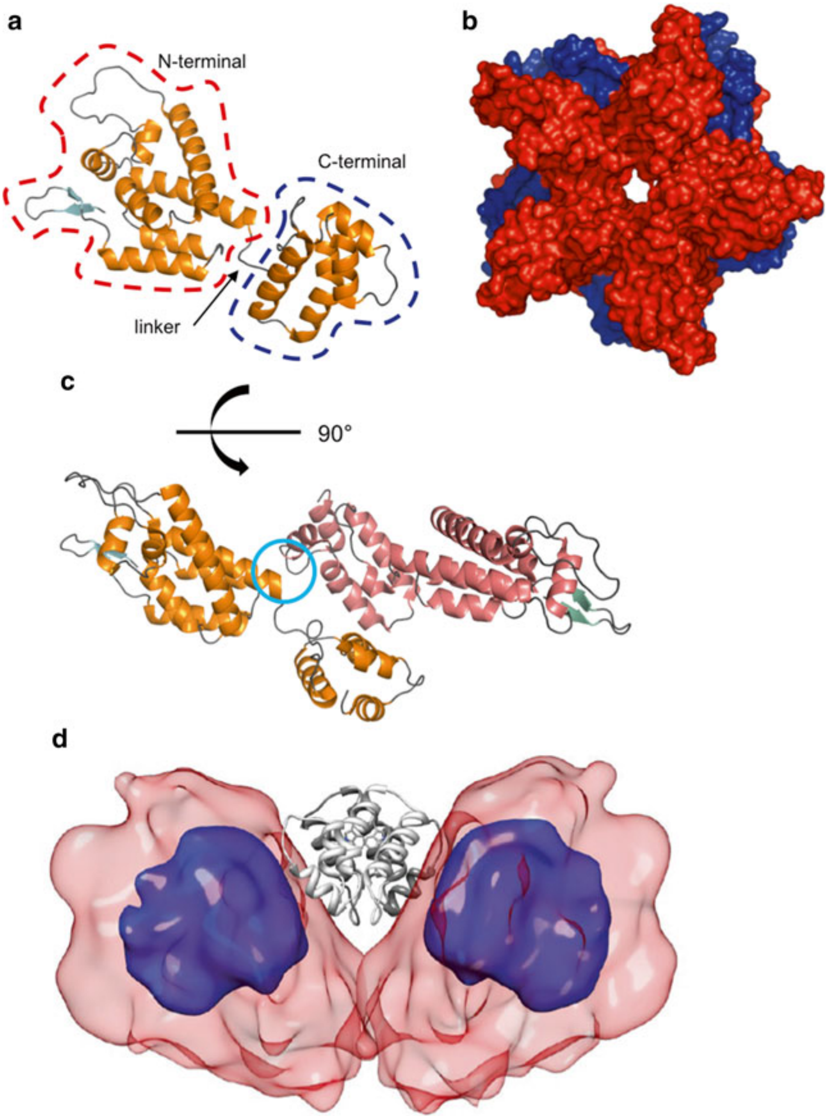


Fig. 22.4 The example of the HIV-1 capsid protein. (**a**) Ribbon representation of the full-length monomer (β-strands, α helices and loops are indicated in *cyan*, *orange* and *grey* respectively) (PDB ID: 2M8N) [57]. The N- and the C-terminal domains and the flexible linker region are highlighted. (**b**) Surface representation of the pentamer (the N-terminal and C-terminal domains are indicated in *red* and *blue* respectively) (PDB ID: 3PO5) [60]. (**c**) Ribbon representation of the full length dimeric protein (β-strands, α helixes and loops are indicated in *cyan*, *orange* and *grey* respectively in Chain A and in *pale green*, *salmon* and *grey* in chain B) (PDB ID: 2M8L) [57]. The contact between the N-terminal domain of chain A and the C-terminal domain of chain B is highlighted. (**d**) Structural ensembles calculated for the full-length monomeric HIV-1 capsid protein. The overall distribution of the N-terminal domain relative to the C-terminal domain (*light* and *dark gray* ribbons) is displayed as a reweighted atomic probability plotted at 50 % (*blue*) and 10 % (*transparent red*) of the maximum value [57] (Panel **d** was reprinted with permission from Deshmukh et al. [57]. Copyright 2013 American Chemical Society)

Worth mentioning are two other similar approaches which use a combination of NMR and SAXS data. Sattler and co-workers [69] implemented in the CNS package an algorithm which performs a topology refinement of a complex using previously solved structures which are refined against SAXS and NMR RDC data. This approach is based on a first step where the radius of gyration of the complex is used to refine inter-domain distances and a second step in which SAXS data at higher angles are used to define domain positions. This procedure was tested on the barnase/barstar complex [69]. The authors developed further this approach, allowing a combination of SAXS data with any type of NMR restraints in a standard structure calculation set-up [70].

The DADIMODO software was initially born to optimize multidomain homology models using RDC and SAXS data [71]. The algorithm was enhanced and extended to allow refinement of proteins and molecular complexes using NMR derived distance and orientational restraints [72]. This program introduces "mutations" (ψ and φ backbone torsion angles are modified by a random amount), and performs an energy minimization to amend backbone distortions. It then selects the conformations that converge in an energy minimum. Survivors of this first step are fitted versus experimental data, selected and used again in the mutation step. The number of cycles is determined by the user. DADIMODO was tested on the human spire protein (a two WH2 domains protein) and a two-domain fragment of the ribosomal S1 protein.

22.5 Multi-subunit Complexes

The combined use of SAXS and NMR was adopted by Parson et al. [73] for the study of TolR. This protein is part of the Pal/Tol system, which forms a five-member, membrane-spanning, multi-protein complex that is involved in several cellular processes (*e.g.*, bacterial outer membrane integrity [74], cell division [75]) and is a potential target for treatment against antibiotic-resistant bacteria. In this study the authors solved the NMR structure of periplasmatic domain of TolR from *Haemophilus influenzae* using conventional NOE restrains. According to gel filtration and light scattering, TolR is a dimer. The protein has a secondary structure βββαβα with β4 pairing up with the other protomer in an antiparallel manner and forming an eight-stranded β-sheet (Fig. 22.5a). The α-helices lie on the same side of the β-sheet, with α2 from each monomer oriented in an antiparallel fashion. The two β-sheet planes of the monomers are twisted by about 74° respect to each other. The authors used only RDC and SAXS data to reconstitute the monomers' orientation in the dimer. Since only a single set of resonances is present in the ^{15}N-^{1}H HSQC, the authors concluded that the TolR dimer must have C_2 symmetry. There are only three possible combinations due to the restriction imposed by this symmetry [76]. If one monomer (A) is fixed at the origin of the RDC alignment tensor frame, the orientation of the other protomer (B) must correspond to the orientation of A rotated by 180° around either of the x, y or z axes of the external frame (Fig. 22.5b). B was then translated along each of these orientations on a 50 Å radius sphere, using a Fibonacci number-based vector grid, generating rings of spherical distribution of B. Arrangements that resulted having a backbone rmsd below 0.25 Å [77] were selected, generating 800 possible dimers. At this point, B was translated towards A generating dimers with a 2.8 Å minimal distance between the two protomers. Further selection was obtained by fitting every dot in panel B using CRYSOL 2.6 [78] against experimental SAXS data (Fig. 22.5c, d). The best resulting model, with a $\chi^2 = 1.127$, has a rmsd of 0.8 Å compared to the TolR dimer solved with conventional NMR methods. This proved that the methodology successfully identified the correct orientation of the two monomers, even if they have a slight translational shift between the two domains. The authors ascribed the success of their studies as compared to previous attempts using SAXS data alone on γS-Crystalline [43] to the better signal-to noise of the SAXS data and to having a C_2 symmetry.

Similarly, the Wang's group developed an algorithm named GASR (Global Architecture derived from SAXS and RDC) that uses RDC

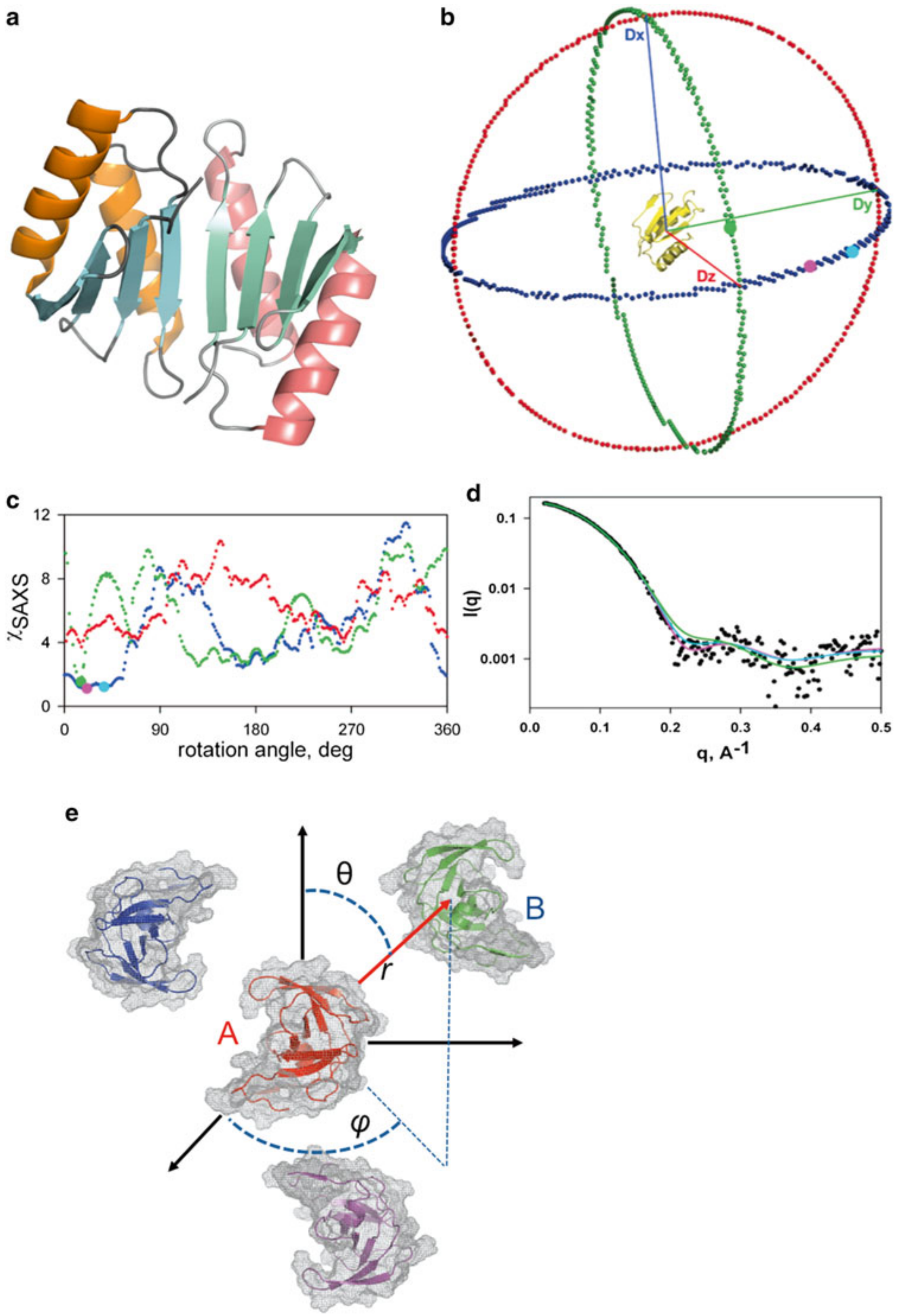


Fig. 22.5 (**a**) Ribbon representation of TolR (PDB ID: 2JWL) [73]. β-strands, α helixes and loops are indicated in *cyan*, *orange* and *grey* respectively in Chain A and in *pale green*, *salmon* and *grey* in chain B. (**b**) Representation of the possible solutions obtained assuming C_2 symmetry. The centers of mass coordinates of domain B are shown as *solid dots*: *blue dots* correspond to the case in which Dx is the C_2 axis (correct solution), *green* and *red dots* correspond to the cases in which the C_2 axis is along the Dy and Dz axes, respectively [73]. (**c**) Plot of the χ values from

and SAXS data to orient subunits and define the global shape of complexes [2]. They benchmarked the software using five different case studies, which included the HIV protease (homodimeric protein), L11 and γD-Crystallin (two two-domain proteins, in which the two domains were treated as two independent interacting proteins), GB1 (weak affinity homodimer), and the ILK ankyrin repeat domain bound to the PINCH LIM1 domain (high affinity dimer). GASR uses a rigid body grid search, conceptually similar to the protocol used by Parson et al. [73] (Fig. 22.5e). Differently from the approach described before, GASR runs two grid searches, the second being a fine search on selected structures. The selection is based on R_g, D_{max}, and D_{min}, which are, respectively, the radius of gyration, the maximum and the minimum linear dimensions between heavy atoms within the two subunits. R_g and D_{max} are easily estimated from experimental SAXS data. D_{min} is set to 1.5 Å for covalently (two domains proteins) linked proteins or to 3.0 Å for transiently interacting proteins, similarly to Parson et al. [73] that used 2.8 Å. Models generated during the second step are analyzed using a probability distribution.

A genuine *de novo* model generated with GASR was recently published by Hirano et al. [14]. In their NMR study of the conformational dynamics of Lys48-linked di-ubiquitin, the authors used GASR to determine the relative orientation and position of the two ubiquitin subunits in a cyclic Lys48-linked di-ubiquitin. The best model generated by the program revealed that the solution structure of cyclic Lys48-linked di-ubiquitin bears a close resemblance to previously reported crystal structures of the non-cyclic counterpart.

Wang et al. [15] applied a procedure similar to that implemented in GASR for the oligomerization study of CCL5. This protein is a pro-inflammatory chemokine, which has a propensity for aggregation and is essential for migration *in vivo*, T cell activation and apoptosis, and HIV entry into cells. Previous structural studies had not explored the quaternary conformation of CCL5 higher order oligomers. Initial analysis of NMR, SAXS and DLS data suggested that CCL5 is mainly a tetramer in solution, with the presence of hexamer species. The model generated in this work was obtained through a simple grid search restrained by the symmetry and shape of the tetramer, using a procedure similar to that described by Wang et al. [2]. Model were selected using SAXS data and a favorable binding surface using a residue-pairing score [16]. Due to the presence of tetramer-hexamer equilibrium, the hexamer model was produced by adding an additional dimer unit to the tetramer, duplicating the initial dimer-dimer interface. The tetramer-hexamer ratio was adjusted using OLIGOMER [79] to find the best fit to the SAXS data. The best model fitted the SAXS data with a $\chi^2 = 1.13$ when using 40 % tetramer and 60 % hexamer. This model forms a tetramer interface which pairs β2 of a protomer in dimer A with the α helix of a protomer in dimer B. NMR cross saturation experiments were used to confirm the inter-dimer interface defined in the grid search.

The study by NMR of RNA-RNA complexes presents several difficulties mainly attributable to their elongated structures. The number of hydrogen atoms in RNA is small as compared to proteins leading to a considerably reduced proton spin density. NOE experiments are in general rather insensitive and lack of signal dispersion complicates resonance assignment. Although different RDC datasets should always be recorded using different alignment methods, different media give rise to similar alignment tensors for

Fig. 22.5 (continued) experimental SAXS data versus the rotation angle. The positions of the three best fitting geometries are shown in *magenta*, *cyan*, and *green* [73]. (**d**) Fit of the three dimer geometries to the experimental scattering data (*black dots*) with the color scheme matching panel **c** [73]. (**e**) Illustration of spatial search used in the GASR program for a two-subunit protein in a spherical polar axis system. Shown here are the two subunits of the HIV-1 protease. Subunit A in *red* is fixed at the axis origin, while subunit B has three discrete possible orientations, depicted in *magenta*, *green* and *blue* "translated" around subunit A without "change in orientation relative to subunit A" [96] (Panels **a–c** were reprinted with permission from Parsons et al. [73]. Copyright 2008 American Chemical Society)

RNA and do not resolve orientation degeneracy [80]. SAXS-aided procedures thus play a major role in the study of RNA-RNA complexes. GASR was tested on a 30 kDa homodimeric tetraloop-receptor RNA complex [81], which is a commonly occurring RNA tertiary structural motif involved in helical packing [82]. This complex structure was previously solved (PDB ID: 2JYJ) using conventional NMR spectroscopy [83, 84]. The rmsd between the best SAXS-defined dimer and the PDB file 2JYJ was found to be 0.4 Å indicating a clear consistency between the two structures. The interaction interfaces were also almost identical including hydrogen bonds and base stacking. The authors stressed the importance of using SAXS data for structural refinement and the substantial difference of the final SAXS-refined model from the model generated without SAXS information, which is much shorter with an rmsd between the two models of 3.2 Å.

22.6 A Case Study: Frataxin and the Iron-Sulfur Cluster Machinery

One of the main ongoing projects in our group is the study of Friedreich's ataxia, a relentless and currently incurable neurodegenerative disease [85]. This disease is caused by a reduced expression level of frataxin, an essential iron-binding protein highly conserved from bacteria to humans. Using the bacterial frataxin ortholog, CyaY, we showed that CyaY participates in iron-sulfur (Fe-S) cluster assembly as an iron-dependent inhibitor of cluster formation, through binding to the desulfurase IscS [86]. We proposed that frataxins are iron sensors that act as regulators of Fe-S cluster formation to fine-tune the quantity of Fe-S cluster formed to the concentration of the available acceptors [86]. This is a highly conserved machine, which ensures the formation of these essential prosthetic groups and their transfer to the final acceptors. Central to the machine are the two components IscS and IscU (using the bacterial name or Nfs1 and Isu in eukaryotes). IscS is a PLP-dependent cysteine desulfurase, which delivers sulfur for Fe-S cluster synthesis to IscU, a Fe-S scaffold protein on which the Fe-S cluster is assembled [87, 88]. To support our enzymology studies with structural evidence, we resolved to model the IscS-IscU complex bound to frataxin using the bacterial proteins. To obtain a molecular description of the IscS-IscU-CyaY complex (where CyaY is the bacterial ortholog of frataxin), we first tried to crystallize the binary complexes obtaining good quality crystals under several different conditions. Unfortunately, they contained only IscS [89]. We thus used an alternative approach based on NMR restrained molecular docking simulations validated by experimental SAXS data. CSP data were used to identify the CyaY and IscU surfaces of interaction with IscS using ^{2}H, ^{15}N double-labeled CyaY (or IscU) and titrating these proteins with IscS (Figs. 22.6a, b). The interaction surface of IscS was defined by titrating ^{2}H, ^{15}N double-labelled CyaY (or IscU) with carefully designed IscS mutants, chosen to target residues that could potentially affect the interaction. The docking software HADDOCK [90], which allows the use of "protein interfaces in ambiguous interaction restraints" (AIRs) to drive the docking process, was used to model the ternary complex. Models were then scored and experimentally verified using SAXS data (Fig. 22.6c). Publication of the first high-resolution structure of the IscS-IscU complex gave us confidence in our procedure [91].

Our model also clarified whether frataxin interacts with IscS or with IscU. The question

Fig. 22.6 (continued) (PDB ID: 1SOY) [96]. Helical and β-sheet regions are indicated in *orange* and *cyan* respectively. The side chains of the residues involved in IscS binding are explicitly shown in *blue*. (**c**) Comparison of the SAXS densities superposed to the crystal structures of IscS (PDB ID: 1P3W) [97] for IscS alone, IscS/IscU, CyaY/IscS and CyaY/IscS/IscU. Regions with additional density in the binary and ternary complexes are highlighted in *red* or *yellow ovals*. (**d**) Final ternary model of the IscS, IscU and CyaY complex. CyaY is shown as a *red ribbon* while the two subunits of IscS homodimer and IscU are shown respectively as *violet*, *pink* and *cyan* molecular surfaces. The side chain of the conserved Trp61 of CyaY is shown in *blue*

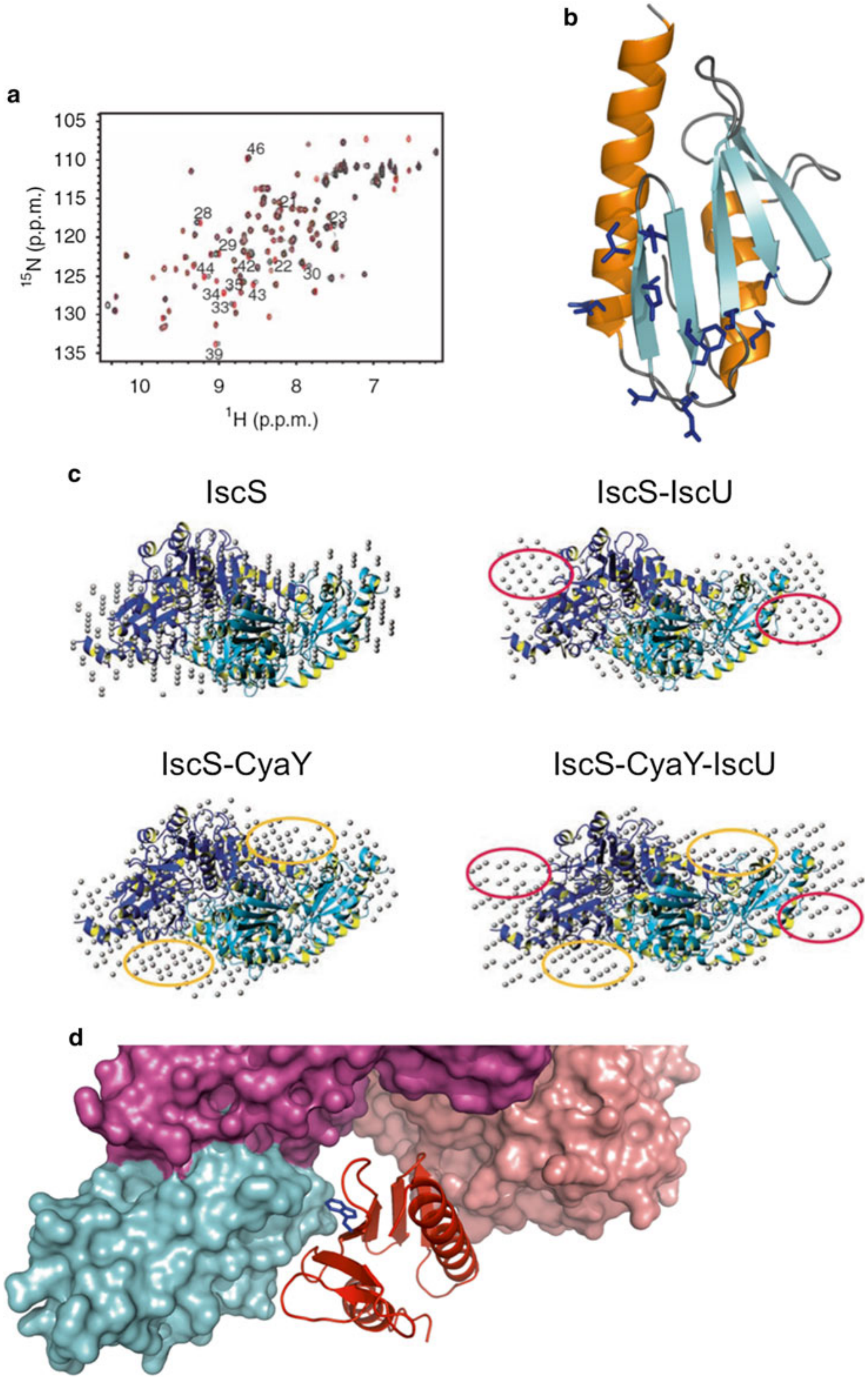


Fig. 22.6 Modeling the ternary complex of IcsS/IscU/CyaY. (**a**) Comparison of the NMR HSQC spectra of ^{15}N labeled CyaY recorded at 25 °C and 800 MHz in the absence (*red*) and in the presence (*black*) of unlabeled IscS (at a protein ratio of 1:0.8). Residues affected by the titration are marked. (**b**) Ribbon representation of the CyaY structure

was raised because while a direct interaction between human frataxin and the eukaryotic ortholog of IscU had been previously reported [92], experiments on the bacterial proteins had proved negative. We showed experimentally that CyaY packs mainly against IscS while limited interactions with IscU are possible but only in the context of the ternary complex [93] (Fig. 22.6d). Interestingly, the contact surface between CyaY and IscU involves Trp61 of CyaY, a highly conserved residue that is known to be indispensable for the binding of human frataxin with Isu [92]. Our model also indicates that formation of the IscS-IscU-CyaY complex does not require the presence of iron, in contrast to previously published data on the yeast frataxin [94]. On the opposite, the surface of interaction involves direct recognition of a highly negatively charged region of CyaY by a positively charged patch on IscS, thus strongly suggesting that an active role of iron in complex formation is unlikely.

More recently, we applied the same procedure to the study of the Fe-S cluster core machinery (IscS-IscU) in the presence of ferredoxin showing that this protein competes with the same binding site previously determined to accommodate CyaY [95].

22.7 Conclusions

It is clear that the future of Structural Biology relies on the combination of different techniques rather than the development of one unique methodology with the hope that this could solve the incredible complexity of Biology. Hybrid methodologies seem to provide a flexible and adaptable answer, which is worth expanding and potentiating. We thus hope that this review contributes to spreading the information and encourages always new groups to develop novel and more powerful approaches to the study of complex systems by NMR.

Acknowledgments AP was funded by MRC (Grant ref. U117584256) and the EFACTS EU consortium.

References

1. Koch MH, Vachette P, Svergun DI (2003) Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. Q Rev Biophys 36(2):147–227
2. Wang J, Zuo X, Yu P, Byeon IJ, Jung J, Wang X, Dyba M, Seifert S, Schwieters CD, Qin J, Gronenborn AM, Wang YX (2009) Determination of multicomponent protein structures in solution using global orientation and shape restraints. J Am Chem Soc 131(30):10507–10515
3. Robinson CV, Sali A, Baumeister W (2007) The molecular sociology of the cell. Nature 450(7172):973–982
4. Ramakrishnan V (2011) Molecular biology. The eukaryotic ribosome. Science 331(6018):681–682
5. Neuhaus D, Williamson MP (2000) The nuclear overhauser effect in structural and conformational analysis, 2nd edn. Wiley, New York
6. Ernst RR, Bodenhausen G, Wokaun A (1997) Principles of nuclear magnetic resonance in one and two dimensions. The international series of monographs on chemistry, vol 14. Clarendon Press; Oxford University Press, Oxford
7. Tugarinov V, Hwang PM, Ollerenshaw JE, Kay LE (2003) Cross-correlated relaxation enhanced 1H[bond]13C NMR spectroscopy of methyl groups in very high molecular weight proteins and protein complexes. J Am Chem Soc 125(34):10420–10428
8. Clore GM, Gronenborn AM (1989) Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy. Crit Rev Biochem Mol Biol 24(5):479–564
9. Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science 278(5340):1111–1114
10. Al-Hashimi HM, Valafar H, Terrell M, Zartler ER, Eidsness MK, Prestegard JH (2000) Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. J Magn Reson 143(2):402–406
11. Clore GM (2000) Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear overhauser enhancement data and dipolar couplings by rigid body minimization. Proc Natl Acad Sci U S A 97(16):9021–9025
12. Rumpel S, Becker S, Zweckstetter M (2008) High-resolution structure determination of the CylR2 homodimer using paramagnetic relaxation enhancement and structure-based prediction of molecular alignment. J Biomol NMR 40(1):1–13
13. Clore GM, Schwieters CD (2003) Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from

1H/15N chemical shift mapping and backbone 15N-1H residual dipolar couplings using conjoined rigid body/torsion angle dynamics. J Am Chem Soc 125(10):2902–2912

14. Hirano T, Serve O, Yagi-Utsumi M, Takemoto E, Hiromoto T, Satoh T, Mizushima T, Kato K (2011) Conformational dynamics of wild-type Lys-48-linked diubiquitin in solution. J Biol Chem 286(43):37496–37502
15. Wang X, Watson C, Sharp JS, Handel TM, Prestegard JH (2011) Oligomeric structure of the chemokine CCL5/RANTES from NMR, MS, and SAXS data. Structure 19(8):1138–1148
16. Moont G, Gabb HA, Sternberg MJ (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 35(3):364–373
17. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature 450(7171):913–916
18. Bernado P, Blackledge M (2010) Structural biology: proteins in dynamic equilibrium. Nature 468(7327):1046–1048
19. Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov MV, Pierattelli R, Ravera E, Svergun DI (2010) Conformational space of flexible biological macromolecules from average data. J Am Chem Soc 132(38):13553–13558
20. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc 129(17):5656–5664
21. Sunnerhagen M, Olah GA, Stenflo J, Forsen S, Drakenberg T, Trewhella J (1996) The relative orientation of Gla and EGF domains in coagulation factor X is altered by Ca2+ binding to the first EGF domain. A combined NMR-small angle X-ray scattering study. Biochemistry 35(36):11547–11559
22. Neylon C (2008) Small angle neutron and X-ray scattering in structural biology: recent examples from the literature. Eur Biophys J 37(5):531–541
23. Improta S, Krueger JK, Gautel M, Atkinson RA, Lefevre JF, Moulton S, Trewhella J, Pastore A (1998) The assembly of immunoglobulin-like modules in titin: implications for muscle elasticity. J Mol Biol 284(3):761–777
24. Banci L, Bertini I, Bren KL, Cremonini MA, Gray HB, Luchinat C, Turano P (1996) The use of pseudocontact shifts to refine solution structures of paramagnetic metalloproteins: Met80Ala cyanocytochrome c as an example. J Biol Inorg Chem 1:117–126
25. Svergun D, Barberato C, Koch MH (1995) CRYSOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 28:768–773
26. Yamniuk AP, Vogel HJ (2004) Calmodulin's flexibility allows for promiscuity in its interactions with target proteins and peptides. Mol Biotechnol 27(1):33–57
27. Chattopadhyaya R, Meador WE, Means AR, Quiocho FA (1992) Calmodulin structure refined at 1.7 A resolution. J Mol Biol 228(4):1177–1192
28. Babu YS, Bugg CE, Cook WJ (1988) Structure of calmodulin refined at 2.2 A resolution. J Mol Biol 204(1):191–204
29. Fallon JL, Quiocho FA (2003) A closed compact structure of native Ca(2+)-calmodulin. Structure 11(10):1303–1307
30. Maximciuc AA, Putkey JA, Shamoo Y, Mackenzie KR (2006) Complex of calmodulin with a ryanodine receptor target reveals a novel, flexible binding mode. Structure 14(10):1547–1556
31. de Diego I, Kuper J, Bakalova N, Kursula P, Wilmanns M (2010) Molecular basis of the death- associated protein kinase-calcium/calmodulin regulator complex. Sci Signal 3(106):ra6
32. Aoyagi M, Arvai AS, Tainer JA, Getzoff ED (2003) Structural basis for endothelial nitric oxide synthase binding to calmodulin. EMBO J 22(4):766–775
33. Meador WE, Means AR, Quiocho FA (1993) Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures. Science 262(5140):1718–1721
34. Meador WE, Means AR, Quiocho FA (1992) Target enzyme recognition by calmodulin: 2.4 A structure of a calmodulin-peptide complex. Science 257(5074):1251–1255
35. Kurokawa H, Osawa M, Kurihara H, Katayama N, Tokumitsu H, Swindells MB, Kainosho M, Ikura M (2001) Target-induced conformational adaptation of calmodulin revealed by the crystal structure of a complex with nematode Ca(2+)/calmodulin-dependent kinase kinase peptide. J Mol Biol 312(1):59–68
36. Chou JJ, Li S, Klee CB, Bax A (2001) Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains. Nat Struct Biol 8(11):990–997
37. Ikura M, Clore GM, Gronenborn AM, Zhu G, Klee CB, Bax A (1992) Solution structure of a calmodulin-target peptide complex by multidimensional NMR. Science 256(5057):632–638
38. Hoeflich KP, Ikura M (2002) Calmodulin in action: diversity in target recognition and activation mechanisms. Cell 108(6):739–742
39. Huang JR, Warner LR, Sanchez C, Gabel F, Madl T, Mackereth CD, Sattler M, Blackledge M (2014) Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. J Am Chem Soc 136(19):7068–7076
40. Mackereth CD, Madl T, Bonnal S, Simon B, Zanier K, Gasch A, Rybin V, Valcarcel J, Sattler M (2011) Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. Nature 475(7356):408–411

41. Salmon L, Nodet G, Ozenne V, Yin G, Jensen MR, Zweckstetter M, Blackledge M (2010) NMR characterization of long-range order in intrinsically disordered proteins. J Am Chem Soc 132(24):8407–8418
42. Guerry P, Salmon L, Mollica L, Ortega Roldan JL, Markwick P, van Nuland NA, McCammon JA, Blackledge M (2013) Mapping the population of protein conformational energy sub-states from NMR dipolar couplings. Angew Chem Int Ed Engl 52(11):3181–3185
43. Grishaev A, Wu J, Trewhella J, Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. J Am Chem Soc 127(47):16621–16628
44. Guo DY, Blessing RH, Langs DA (2000) Globbic approximation in low-resolution direct-methods phasing. Acta Crystallogr D Biol Crystallogr 56(Pt 9):1148–1155
45. Svergun DI, Petoukhov MV, Koch MH (2001) Determination of domain structure of proteins from X-ray solution scattering. Biophys J 80(6):2946–2953
46. Chacon P, Moran F, Diaz JF, Pantos E, Andreu JM (1998) Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm. Biophys J 74(6):2760–2775
47. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54(Pt 5):905–921
48. Grishaev A, Tugarinov V, Kay LE, Trewhella J, Bax A (2008) Refined solution structure of the 82- kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. J Biomol NMR 40(2):95–106
49. Grishaev A, Ying J, Canny MD, Pardi A, Bax A (2008) Solution structure of tRNAVal from refinement of homology model against residual dipolar coupling and SAXS data. J Biomol NMR 42(2):99–109
50. Schwieters CD, Suh JY, Grishaev A, Ghirlando R, Takayama Y, Clore GM (2010) Solution structure of the 128 kDa enzyme I dimer from Escherichia coli and its 146 kDa complex with HPr using residual dipolar couplings and small- and wide-angle X-ray scattering. J Am Chem Soc 132(37):13026–13045
51. Takayama Y, Schwieters CD, Grishaev A, Ghirlando R, Clore GM (2011) Combined use of residual dipolar couplings and solution X-ray scattering to rapidly probe rigid-body conformational transitions in a non-phosphorylatable active-site mutant of the 128 kDa enzyme I dimer. J Am Chem Soc 133(3):424–427
52. Kumaraswamy VS, Lindley PF, Slingsby C, Glover ID (1996) An eye lens protein-water structure: 1.2 A resolution structure of gammaB-crystallin at 150 K. Acta Crystallogr D Biol Crystallogr 52(Pt 4):611–622
53. Basak A, Bateman O, Slingsby C, Pande A, Asherie N, Ogun O, Benedek GB, Pande J (2003) High-resolution X-ray crystal structures of human gammaD crystallin (1.25 A) and the R58H mutant (1.15 A) associated with aculeiform cataract. J Mol Biol 328(5):1137–1147
54. Smith CV, Huang CC, Miczak A, Russell DG, Sacchettini JC, Honer zu Bentrup K (2003) Biochemical and structural studies of malate synthase from Mycobacterium tuberculosis. J Biol Chem 278(3):1735–1743
55. Howard BR, Endrizzi JA, Remington SJ (2000) Crystal structure of Escherichia coli malate synthase G complexed with magnesium and glyoxylate at 2.0 A resolution: mechanistic implications. Biochemistry 39(11):3156–3168
56. Tugarinov V, Choy WY, Orekhov VY, Kay LE (2005) Solution NMR-derived global fold of a monomeric 82-kDa enzyme. Proc Natl Acad Sci U S A 102(3):622–627
57. Deshmukh L, Schwieters CD, Grishaev A, Ghirlando R, Baber JL, Clore GM (2013) Structure and dynamics of full-length HIV-1 capsid protein in solution. J Am Chem Soc 135(43):16133–16147
58. Ganser BK, Li S, Klishko VY, Finch JT, Sundquist WI (1999) Assembly and analysis of conical models for the HIV-1 core. Science 283(5398):80–83
59. Ganser-Pornillos BK, von Schwedler UK, Stray KM, Aiken C, Sundquist WI (2004) Assembly properties of the human immunodeficiency virus type 1 CA protein. J Virol 78(5):2545–2552
60. Pornillos O, Ganser-Pornillos BK, Yeager M (2011) Atomic-level modelling of the HIV capsid. Nature 469(7330):424–427
61. Borsetti A, Ohagen A, Gottlinger HG (1998) The C-terminal half of the human immunodeficiency virus type 1 Gag precursor is sufficient for efficient particle assembly. J Virol 72(11):9313–9317
62. Accola MA, Strack B, Gottlinger HG (2000) Efficient particle production by minimal Gag constructs which retain the carboxy-terminal domain of human immunodeficiency virus type 1 capsid- p2 and a late assembly domain. J Virol 74(12):5395–5402
63. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160(1):65–73
64. Byeon IJ, Meng X, Jung J, Zhao G, Yang R, Ahn J, Shi J, Concel J, Aiken C, Zhang P, Gronenborn AM (2009) Structural convergence between Cryo-EM and NMR reveals intersubunit interactions critical for HIV-1 capsid function. Cell 139(4):780–790
65. Gamble TR, Yoo S, Vajdos FF, von Schwedler UK, Worthylake DK, Wang H, McCutcheon JP, Sundquist WI, Hill CP (1997) Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein. Science 278(5339):849–853
66. Worthylake DK, Wang H, Yoo S, Sundquist WI, Hill CP (1999) Structures of the HIV-1 capsid protein dimerization domain at 2.6 A resolution. Acta Crystallogr D Biol Crystallogr 55(Pt 1):85–92

67. Kelly BN, Kyere S, Kinde I, Tang C, Howard BR, Robinson H, Sundquist WI, Summers MF, Hill CP (2007) Structure of the antiviral assembly inhibitor CAP-1 complex with the HIV-1 CA protein. J Mol Biol 373(2):355–366
68. Lemke CT, Titolo S, von Schwedler U, Goudreau N, Mercier JF, Wardrop E, Faucher AM, Coulombe R, Banik SS, Fader L, Gagnon A, Kawai SH, Rancourt J, Tremblay M, Yoakim C, Simoneau B, Archambault J, Sundquist WI, Mason SW (2012) Distinct effects of two HIV-1 capsid assembly inhibitor families that bind the same site within the N-terminal domain of the viral CA protein. J Virol 86(12):6643–6655
69. Gabel F, Simon B, Sattler M (2006) A target function for quaternary structural refinement from small angle scattering and NMR orientational restraints. Eur Biophys J 35(4):313–327
70. Gabel F, Simon B, Nilges M, Petoukhov M, Svergun D, Sattler M (2008) A structure refinement protocol combining NMR residual dipolar couplings and small angle scattering restraints. J Biomol NMR 41(4):199–208
71. Mareuil F, Sizun C, Perez J, Schoenauer M, Lallemand JY, Bontems F (2007) A simple genetic algorithm for the optimization of multidomain protein homology models driven by NMR residual dipolar coupling and small angle X-ray scattering data. Eur Biophys J 37(1):95–104
72. Evrard G, Mareuil F, Bontems F, Sizun C, Perez J (2011) DADIMODO: a program for refining the structure of multidomain proteins and complexes against small-angle scattering data and NMR-derived restraints. J Appl Crystallogr 44:7
73. Parsons LM, Grishaev A, Bax A (2008) The periplasmic domain of TolR from Haemophilus influenzae forms a dimer with a large hydrophobic groove: NMR solution structure and comparison to SAXS data. Biochemistry 47(10):3131–3142
74. Cascales E, Bernadac A, Gavioli M, Lazzaroni JC, Lloubes R (2002) Pal lipoprotein of Escherichia coli plays a major role in outer membrane integrity. J Bacteriol 184(3):754–759
75. Gerding MA, Ogata Y, Pecora ND, Niki H, de Boer PA (2007) The trans-envelope Tol-Pal complex is part of the cell division machinery and required for proper outer-membrane invagination during cell constriction in E. coli. Mol Microbiol 63(4):1008–1025
76. Al-Hashimi HM, Bolon PJ, Prestegard JH (2000) Molecular symmetry as an aid to geometry determination in ligand protein complexes. J Magn Reson 142(1):153–158
77. Dam J, Baber J, Grishaev A, Malchiodi EL, Schuck P, Bax A, Mariuzza RA (2006) Variable dimerization of the Ly49A natural killer cell receptor results in differential engagement of its MHC class I ligand. J Mol Biol 362(1):102–113
78. Konarev PV, Petoukhov MV, Volkov VV, Svergun DI (2006) ATSAS 2.1, a program package for small-angle scattering data analysis. J Appl Crystallogr 39:277–286
79. Latham MP, Hanson P, Brown DJ, Pardi A (2008) Comparison of alignment tensors generated for native tRNA(Val) using magnetic fields and liquid crystalline media. J Biomol NMR 40(2):83–94
80. Zuo X, Wang J, Foster TR, Schwieters CD, Tiede DM, Butcher SE, Wang YX (2008) Global molecular structure and interfaces: refining an RNA:RNA complex structure using solution X-ray scattering data. J Am Chem Soc 130(11):3292–3293
81. Costa M, Michel F (1995) Frequent use of the same tertiary motif by self-folding RNAs. EMBO J 14(6):1276–1285
82. Davis JH, Tonelli M, Scott LG, Jaeger L, Williamson JR, Butcher SE (2005) RNA helical packing in solution: NMR structure of a 30 kDa GAAA tetraloop-receptor complex. J Mol Biol 351(2):371–382
83. Davis JH, Foster TR, Tonelli M, Butcher SE (2007) Role of metal ions in the tetraloop-receptor complex as analyzed by NMR. RNA 13(1):76–86
84. Pandolfo M, Pastore A (2009) The pathogenesis of Friedreich ataxia and the structure and function of frataxin. J Neurol 256(Suppl 1):9–17
85. Adinolfi S, Iannuzzi C, Prischi F, Pastore C, Iametti S, Martin SR, Bonomi F, Pastore A (2009) Bacterial frataxin CyaY is the gatekeeper of iron-sulfur cluster formation catalyzed by IscS. Nat Struct Mol Biol 16(4):390–396
86. Urbina HD, Silberg JJ, Hoff KG, Vickery LE (2001) Transfer of sulfur from IscS to IscU during Fe/S cluster assembly. J Biol Chem 276(48):44521–44526
87. Adinolfi S, Rizzo F, Masino L, Nair M, Martin SR, Pastore A, Temussi PA (2004) Bacterial IscU is a well folded and functional single domain protein. Eur J Biochem 271(11):2093–2100
88. Prischi F, Pastore C, Carroni M, Iannuzzi C, Adinolfi S, Temussi P, Pastore A (2010) Of the vulnerability of orphan complex proteins: the case study of the E. coli IscU and IscS proteins. Protein Expr Purif 73(2):161–166
89. Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125(7):1731–1737
90. Shi R, Proteau A, Villarroya M, Moukadiri I, Zhang L, Trempe JF, Matte A, Armengod ME, Cygler M (2010) Structural basis for Fe-S cluster assembly and tRNA thiolation mediated by IscS protein-protein interactions. PLoS Biol 8(4), e1000354
91. Yoon T, Cowan JA (2003) Iron-sulfur cluster biosynthesis. Characterization of frataxin as an iron donor for assembly of [2Fe-2S] clusters in ISU-type proteins. J Am Chem Soc 125(20):6078–6084
92. Prischi F, Konarev PV, Iannuzzi C, Pastore C, Adinolfi S, Martin SR, Svergun DI, Pastore A (2010) Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly. Nat Commun 1:95

93. Li H, Gakh O, Smith DY, Isaya G (2009) Oligomeric yeast frataxin drives assembly of core machinery for mitochondrial iron-sulfur cluster synthesis. J Biol Chem 284(33):21971–21980

94. Yan R, Konarev PV, Iannuzzi C, Adinolfi S, Roche B, Kelly G, Simon L, Martin SR, Py B, Barras F, Svergun DI, Pastore A (2013) Ferredoxin competes with bacterial frataxin in binding to the desulfurase IscS. J Biol Chem 288(34):24777–24787

95. Yamazaki T, Hinck AP, Wang YX, Nicholson LK, Torchia DA, Wingfield P, Stahl SJ, Kaufman JD, Chang CH, Domaille PJ, Lam PY (1996) Three-dimensional solution structure of the HIV-1 protease complexed with DMP323, a novel cyclic urea-type inhibitor, determined by nuclear magnetic resonance spectroscopy. Protein Sci 5(3):495–506

96. Nair M, Adinolfi S, Pastore C, Kelly G, Temussi P, Pastore A (2004) Solution structure of the bacterial frataxin ortholog, CyaY: mapping the iron binding sites. Structure 12(11):2037–2048

97. Cupp-Vickery JR, Urbina H, Vickery LE (2003) Crystal structure of IscS, a cysteine desulfurase from Escherichia coli. J Mol Biol 330(5):1049–1059

Index

M.C. Vega (ed.), *Advanced Technologies for Protein Complex Production and Characterization*, Advances in Experimental Medicine and Biology 896, DOI 10.1007/978-3-319-27216-0