Springer Protocols

Terry J. McGenity
Kenneth N. Timmis
Balbina Nogales *Editors*

Hydrocarbon and Lipid Microbiology Protocols

Synthetic and Systems Biology – Tools

Springer

Springer Protocols Handbooks

More information about this series at http://www.springer.com/series/8623

Terry J. McGenity · Kenneth N. Timmis · Balbina Nogales
Editors

Hydrocarbon and Lipid Microbiology Protocols

Synthetic and Systems Biology - Tools

Springer

Editors
Terry J. McGenity
School of Biological Sciences
University of Essex
Colchester, Essex, UK

Kenneth N. Timmis
Institute of Microbiology
Technical University Braunschweig
Braunschweig, Germany

Balbina Nogales
Department of Biology
University of the Balearic Islands
and Mediterranean Institute
for Advanced Studies
(IMEDEA, UIB-CSIC)
Palma de Mallorca, Spain

ISSN 1949-2448 ISSN 1949-2456 (electronic)
Springer Protocols Handbooks
ISBN 978-3-662-50430-7 ISBN 978-3-662-50432-1 (eBook)
DOI 10.1007/978-3-662-50432-1

Library of Congress Control Number: 2016938230

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Berlin Heidelberg

Preface to Hydrocarbon and Lipid Microbiology Protocols[1]

All active cellular systems require water as the principal medium and solvent for their metabolic and ecophysiological activities. Hydrophobic compounds and structures, which tend to exclude water, although providing *inter alia* excellent sources of energy and a means of biological compartmentalization, present problems of cellular handling, poor bioavailability and, in some cases, toxicity. Microbes both synthesize and exploit a vast range of hydrophobic organics, which includes biogenic lipids, oils and volatile compounds, geochemically transformed organics of biological origin (i.e. petroleum and other fossil hydrocarbons) and manufactured industrial organics. The underlying interactions between microbes and hydrophobic compounds have major consequences not only for the lifestyles of the microbes involved but also for biogeochemistry, climate change, environmental pollution, human health and a range of biotechnological applications. The significance of this "greasy microbiology" is reflected in both the scale and breadth of research on the various aspects of the topic. Despite this, there was, as far as we know, no treatise available that covers the subject. In an attempt to capture the essence of greasy microbiology, the *Handbook of Hydrocarbon and Lipid Microbiology* (http://www.springer.com/life+sciences/microbiology/book/978-3-540-77584-3) was published by Springer in 2010 (Timmis 2010). This five-volume handbook is, we believe, unique and of considerable service to the community and its research endeavours, as evidenced by the large number of chapter downloads. Volume 5 of the handbook, unlike volumes 1–4 which summarize current knowledge on hydrocarbon microbiology, consists of a collection of experimental protocols and appendices pertinent to research on the topic.

A second edition of the handbook is now in preparation and a decision was taken to split off the methods section and publish it separately as part of the Springer Protocols program (http://www.springerprotocols.com/). The multi-volume work *Hydrocarbon and Lipid Microbiology Protocols*, while rooted in Volume 5 of the Handbook, has evolved significantly, in terms of range of topics, conceptual structure and protocol format. Research methods, as well as instrumentation and strategic approaches to problems and analyses, are evolving at an unprecedented pace, which can be bewildering for newcomers to the field and to experienced researchers desiring to take new approaches to problems. In attempting to be comprehensive – a one-stop source of protocols for research in greasy microbiology – the protocol volumes inevitably contain both subject-specific and more generic protocols, including sampling in the field, chemical analyses, detection of specific functional groups of microorganisms and community composition, isolation and cultivation of such organisms, biochemical analyses and activity measurements, ultrastructure and imaging methods, genetic and genomic analyses, systems and synthetic biology tool usage, diverse applications, and

[1] Adapted in part from the Preface to *Handbook of Hydrocarbon and Lipid Microbiology*.

the exploitation of bioinformatic, statistical and modelling tools. Thus, while the work is aimed at researchers working on the microbiology of hydrocarbons, lipids and other hydrophobic organics, much of it will be equally applicable to research in environmental microbiology and, indeed, microbiology in general. This, we believe, is a significant strength of these volumes.

We are extremely grateful to the members of our Scientific Advisory Board, who have made invaluable suggestions of topics and authors, as well as contributing protocols themselves, and to generous *ad hoc* advisors like Wei Huang, Manfred Auer and Lars Blank. We also express our appreciation of Jutta Lindenborn of Springer who steered this work with professionalism, patience and good humour.

Colchester, Essex, UK — Terry J. McGenity
Braunschweig, Germany — Kenneth N. Timmis
Palma de Mallorca, Spain — Balbina Nogales

Reference

Timmis KN (ed) (2010) Handbook of hydrocarbon and lipid microbiology. Springer, Berlin, Heidelberg

Contents

About the Editors

Terry J. McGenity is a Reader at the University of Essex, UK. His Ph.D., investigating the microbial ecology of ancient salt deposits (University of Leicester), was followed by postdoctoral positions at the Japan Marine Science and Technology Centre (JAMSTEC, Yokosuka) and the Postgraduate Research Institute for Sedimentology (University of Reading). His overarching research interest is to understand how microbial communities function and interact to influence major biogeochemical processes. He worked as a postdoc with Ken Timmis at the University of Essex, where he was inspired to investigate microbial interactions with hydrocarbons at multiple scales, from communities to cells, and as both a source of food and stress. He has broad interests in microbial ecology and diversity, particularly with respect to carbon cycling (especially the second most abundantly produced hydrocarbon in the atmosphere, isoprene), and is driven to better understand how microbes cope with, or flourish in hypersaline, desiccated and poly-extreme environments.

Kenneth N. Timmis read microbiology and obtained his Ph.D. at Bristol University, where he became fascinated with the topics of environmental microbiology and microbial pathogenesis, and their interface pathogen ecology. He undertook postdoctoral training at the Ruhr-University Bochum with Uli Winkler, Yale with Don Marvin, and Stanford with Stan Cohen, at the latter two institutions as a Fellow of the Helen Hay Whitney Foundation, where he acquired the tools and strategies of genetic approaches to investigate mechanisms and causal relationships underlying microbial activities. He was subsequently appointed Head of an Independent Research Group at the Max Planck Institute for Molecular Genetics in Berlin, then Professor of Biochemistry in the University of Geneva Faculty of Medicine. Thereafter, he became Director of the Division of Microbiology at the National Research Centre for Biotechnology (GBF)/now the Helmholtz Centre for Infection Research (HZI) and Professor of Microbiology at the Technical University Braunschweig. His group has worked for many years, *inter alia*, on the biodegradation of oil hydrocarbons, especially the genetics and regulation of toluene degradation, pioneered the genetic design and experimental evolution of novel catabolic activities, discovered the new group of marine hydrocarbonoclastic bacteria, and conducted early genome sequencing of bacteria that

became paradigms of microbes that degrade organic compounds (*Pseudomonas putida* and *Alcanivorax borkumensis*). He has had the privilege and pleasure of working with and learning from some of the most talented young scientists in environmental microbiology, a considerable number of which are contributing authors to this series, and in particular Balbina and Terry. He is Fellow of the Royal Society, Member of the EMBO, Recipient of the Erwin Schrödinger Prize, and Fellow of the American Academy of Microbiology and the European Academy of Microbiology. He founded the journals *Environmental Microbiology*, *Environmental Microbiology Reports* and *Microbial Biotechnology*. Kenneth Timmis is currently Emeritus Professor in the Institute of Microbiology at the Technical University of Braunschweig.

Balbina Nogales is a Lecturer at the University of the Balearic Islands, Spain. Her Ph.D. at the Autonomous University of Barcelona (Spain) investigated antagonistic relationships in anoxygenic sulphur photosynthetic bacteria. This was followed by postdoctoral positions in the research groups of Ken Timmis at the German National Biotechnology Institute (GBF, Braunschweig, Germany) and the University of Essex, where she joined Terry McGenity as postdoctoral scientist. During that time, she worked in different research projects on community diversity analysis of polluted environments. After moving to her current position, her research is focused on understanding microbial communities in chronically hydrocarbon-polluted marine environments, and elucidating the role in the degradation of hydrocarbons of certain groups of marine bacteria not recognized as typical degraders.

Systems and Synthetic Biology in Hydrocarbon Microbiology: Tools

Víctor de Lorenzo

Abstract

Systems and synthetic biology represent the two sides of the recent ambition to understand quantitatively biological systems as full, logically organized objects able to perform functions on the basis of their extant blueprint and therefore amenable to being refactored to generate new-to-nature properties. The *systemic* approach focuses on the cataloguing of all components of the studied entity, their relational logic and their dynamic interplay for comprehending and predicting its behaviour as a whole. The *synthetic* counterpart adopts straight engineering principles taken from industrial and electric manufacturing for re-creating biological systems from perfectly defined constituents as well as for constructing functionalities that have not yet emerged through the natural evolutionary course.

Keywords: Containment, DNA assembly, Modelling, Networks, Orthogonality, Parts

1 Systems and Synthetic Biology Meet Environmental Biotechnology

The biological world is an extraordinary case of multi-scale complexity in which every layer of the system (from single genes or proteins up to whole landscapes) seems to be connected both upstream and downstream to other autonomous layers of intricacy and interdependence [1]. The challenge in this scenario is that understanding the rules that govern one level of functioning (say, the way one biodegradation pathway is regulated in a single bacterium in a Petri dish) may tell us very little on the next layer (e.g. whether the same strain/pathway has any significance in degrading the same compound in a natural niche). The phenomenon so pervasive in complex systems (and therefore in biology) known as *emergence* means that the combination of discrete components of a system may not result in a clear fusion of their properties or their parameters but on different qualities that can be better, worse or altogether different of what was there as the starting point. Although molecular biology was born after the WWII out of the interest of physicists for living objects, there was a sort of

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 1–7,
DOI 10.1007/8623_2015_185, © Springer-Verlag Berlin Heidelberg 2015, Published online: 19 March 2016

foundational choice for reductionist approaches that advocated the focus on the details of specific biological constituents as a way to understand the whole. Despite the spectacular development of molecular biology for more than 30 years, the onset of techniques for easy and cheap DNA sequencing and the generation of large volumes of *omics* data on specific microorganisms have both exposed the limitations of such reductionist strategies and opened a possibility to study biological systems as a whole and not as a sum of parts [2]. Ultimately, systemic approaches mean moving from a focus on the separate components of a system towards the whole of dynamic interactions between them. As Henrik Kacser, the founder of metabolic control analysis, stated in a celebrated quote "One thing is certain: if you want to understand the whole you must study the whole" [3]. This criterium can be applied to nearly every level of biological complexity but is particularly relevant to examine environmental microorganisms [2]. In this context, there is a need to incorporate conceptual methods imported from other fields, in particular, the abstractions that are typical of physics and mathematics and the modelling tools often taken from electric or chemical engineering, as well as some principles of complexity theory for tackling the behaviour of, e.g. non-linear systems [4].

Yet, there are some features of biological entities, specifically microorganisms, that have one remarkable property when placed in a multi-scale multifaceted scenario, namely, that the functionalities encoded in their DNA can quickly penetrate through all complexity layers, from the genome to complete landscapes. One dramatic example is the global spread of antibiotic resistance genes from their environmental point of origin, often resulting from minor mutations in metabolic genes towards virtually all hospitals of the world [5]. The unstoppable flow of DNA through all physical and geographical barriers places a considerable focus of contemporary systemic environmental microbiology on the genomes of the bacteria at stake (whether culturable or not) and on attempts to extract from them as much information as possible [2]. This makes a combination of wet data, collecting methods with bioinformatic analysis and modelling necessary, as addressed separately in the chapters below.

The diversity of assets available for systemic and synthetic approaches to environmental microorganisms and their interface with hydrocarbons has distinct but still intertwined aspects. On the one hand, we have the deconstruction-reconstruction-redesign agenda (the last with different qualities of the original system). On the other hand, there are the *wet* vs the in silico strategies. The papers contained in this volume provide a good panel of current protocols that allow addressing both old and new questions under systemic and synthetic perspectives. Note, however, that the methods included in this collection focus exclusively in the aspects that have to do with fundamental *understanding* of the system at stake, not with practical bioengineering. The group of articles

under the umbrella of *synthetic biology* therefore deals with using *synthesis* and directed design as the counterpart of *analysis*, the two being the pillars of any research endeavour. Although some of these concepts and methods described here are equally usable for explicit bioengineering, e.g. for biotechnological purposes, the protocols dealing with applied projections of synthetic biology have been compiled in a separate Volume of this Protocol series (Hydrocarbon and Lipid Microbiology Protocols: Synthetic and Systems Biology Applications).

2 Systemic Approaches to Environmentally Relevant Bacteria

It is generally accepted that the key appeal of systems biology relies on its power to translate loads of data into workable models that help in understanding otherwise complex and undecipherable biological entities. And, as a consequence, they guide experiments for validation and further model refinement. In this respect, systems biology borrows from engineering the typical modelling/testing/improvement cycle that is so typical of industrial design [4]. The part of the cycle that has to do with massive data collection (whether transcriptomics, metabolomics, proteomics, etc.) has been tackled in a different Volume (Hydrocarbon and Lipid Microbiology Protocols: Genetic, Genomic and Systems Analyses of Pure Cultures) and will not be revisited here. The starting point of the chapters below is instead the availability of enough data on the microbial system under scrutiny that allows us to abstract the components and apply models for their comprehension. And as mentioned above, the methodologies discussed range from purely wet to purely computational. One exemplary case is the set-up of metabolic models out of genomic information, as discussed by Nogales [6]. One reliable test of robustness of such in silico models (as well as an awesome source of extra information) is the passing of the strain under study through a large panel of growth conditions (the so-called phenomic testing [7]). The effect of given perturbations (e.g. mutations) on the resulting biochemical network can thereby be checked by observing growth or lack of it on specific carbon sources or stress settings. And once one has a good model in hand, there is a large number of operations that one can perform in vivo to understand specific steps of the metabolic traffic and, wherever desired, modify the metabolism at user's will. Along this line, article [8] showcases one example in which systemic understanding of NAD(P)/NAD(P)H balance in environmental bacteria results in strategies to make it a better host for knocked-in redox reactions. In a further turn of the screw, one can also benefit from modelling and computer-assisted genetic design for engineering new biochemical and catalytic capabilities in pre-existing metabolic networks, in fact one of the most spectacular products of contemporary systems biology [9, 10]. But one important detail is often

overlooked in systems-guided metabolic engineering: reactions do not occur in dimensionless space but in a reactor (the cell) that has a distinct 3D distribution of its material constituents. Biotransformations carried out by whole-cell catalysts must occur in time and space as if in a chemical factory. How are microorganisms inside organized thereof? To answer this question, one thinks immediately in looking at cells directly with a microscope: in fact, super-resolution microscopy is a growingly feasible approach [11, 12]. But one can also generate a considerable amount of information on the same matter through genetic and computational probing of co-occurrent genomic spots, e.g. those that share the same transcriptional factors. As presented in [13], such analyses allow dissection of higher-level genomic architecture and perhaps identification of optimal sites for knocking in new activities. The techniques just mentioned can then be applied in various ways to decipher details of the interplay between specific microorganisms and hydrocarbons as well as gaining insight on the general properties of given bacterial groups specialized in key environmental activities (e.g. methylotrophs [14]).

3 The Tools of Synthetic Biology

In its most widespread connotation, synthetic biology is associated to the deep genetic design of biological systems for given biotechnological purposes (e.g. a sort of extreme genetic engineering [15]). In reality, there is much more than that: synthetic biology adopts engineering not as an analogy or a metaphor but as a veritable interpretative frame [16, 17]. This choice becomes instrumental both for making sense of the relational logic of extant biological objects and for (genetically) changing that logic in order to create new-to-nature activities. Such an agenda (for the sake of both *understanding* and *doing*) involves standardization, metrology, definition of system's boundaries, scalability, etc.; all of them are engineering matters that have been traditionally put aside in biological research. But the benefits of such a take (that involves a new jargon for describing biological properties and transactions as well [16, 17]) start just to be appreciated and expanding at the time of writing this article. For now, synthetic biology approaches have been mostly applied to *E. coli*, *mycoplasma* and yeast and to a much lesser extent to other microorganisms. But the range of environmental bacteria that can be the subject of these approaches is growing. Some of the articles found in this volume map precisely in this ongoing momentum.

Despite the ease and lowering of the price of chemically synthesized DNA, the synthetic biology practitioner still faces the problem of combining many variants of the same biological parts for optimizing a given pathway or construct. It cannot come as a surprise that techniques for the assembly of multiple DNA

segments are still quite a challenge and the subject of considerable efforts. A number of strategies have been proposed to this end, and the article Nørholm [18] describes in detail one that looks specially promising for complex combinations or DNA fragments. A separate matter is that of measuring the activity of the resulting constructs. As mentioned above, metrology is at the core of any serious effort of engineering biological systems but still a phenomenal challenge. Given the context dependency of the parameters associated to virtually all biological functions [19], even the very simple measurement of promoter activity is the subject of controversies [20]. To alleviate this problem, [21] proposes a virtually automated high-throughput method of prototyping promoter activity that can be interfaced with computational analyses in order to unveil the general rules that determine transcriptional activity. This is an important step towards standardization of promoter strength and the description of its potency in polymerase per second (PoPs) units, a biological counterpart of electric current [22].

In addition to DNA assembly and standardization/metrology of the gene expression flow, a third recurrent theme in synthetic biology is orthogonality, e.g. the pursuit of a minimal dependence of the engineered modules or devices on the host and vice versa [23]. Although nature has already produced a large number of parts and devices that work in a fashion somewhat autonomous of the carrier (e.g. activities found in bacteriophages and mobile genetic systems [24]), synthetic biology pushes these natural limits in new directions. As shown in [25], one can reassign the meaning of given triplets of the genetic code to be read by equally modified ribosomal constituents in order to generate orthogonal gene expression devices. These amplify extant biological diversity as they allow incorporation of non-natural or unusual amino acids into the structure of otherwise natural proteins.

4 Outlook

The wealth of conceptual and material tools delivered by contemporary systems and synthetic biology (of which the protocols of this volume are just a sample) is bound to change the way we understand and reshape biological entities as engineer-able objects. At the same time, the creation in the laboratory of new biological activities and the chances of their accidental scape or deliberate release into the environment are not devoid of controversies that echo former debates on the safety of genetically engineered microorganisms [26]. In this context, article [27] reviews updated views on this matter as well as fresh propositions to limit the spread of deeply engineered (or even altogether synthetic) bacteria in a time in which the hype and the promise of synthetic biology are also accompanied by public fears on its possible damage if released out of control.

In sum, incomplete as it is because of the very fast developments in the field, this volume displays a representative palette of quantitative strategies offered by systems and synthetic biology to examine complex (micro)biological scenarios.

Acknowledgements

The work in Author's Laboratory is supported by the CAMBIOS Project of the Spanish Ministry of Economy and Competitiveness, the ARISYS, EVOPROG and EMPOWERPUTIDA Contracts of the EU, the ERANET-IB and the PROMT Project of the CAM.

References

1. de Lorenzo V (2008) Systems biology approaches to bioremediation. Curr Opin Biotechnol 19:579–589
2. de Lorenzo V, Fraile S, Jiménez J (2010) Emerging systems and synthetic biology approaches to hydrocarbon biotechnology. In: Timmis KN, McGenity TJ, van der Meer JR, de Lorenzo V (eds) Handbook of hydrocarbon and lipid microbiology. Springer, pp 1411–1435
3. Kacser H (1986) On parts and wholes in metabolism. In: Welch GR, Clegg JS (eds) The organization of cell metabolism. Plenum, New York, pp 327–337
4. Alon U (2006) An introduction to systems biology: design principles of biological circuits. Chapman and Hall/CRC, Boca Raton
5. Davies J, Davies D (2010) Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev 74:417–433
6. Nogales J (2016) A practical protocol for genome-scale metabolic reconstruction. Springer Protocols Handbooks. doi:10.1007/8623_2015_98
7. Hausler S (2016) Phenome-ing microbes. Springer Protocols Handbooks. doi:10.1007/8623_2015_178
8. Chavarria M (2016) Quantitative physiology approaches to understand and optimize reducing power availability in environmental bacteria. Springer Protocols Handbooks. doi:10.1007/8623_2015_84
9. O'Brien EJ, Monk JM, Palsson BO (2015) Using genome-scale models to predict biological capabilities. Cell 161:971–987
10. Palsson BO (2015) Systems biology. Cambridge University Press, Cambridge
11. Campos M, Jacobs-Wagner C (2013) Cellular organization of the transfer of genetic information. Curr Opin Microbiol 16:171–176
12. Parry BR, Surovtsev IV, Cabeen MT, O'Hern CS, Dufresne ER, Jacobs-Wagner C (2014) The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity. Cell 156:183–194
13. Bouyioukos C, Elati M, Kepes F (2016) Protocols for probing genome architecture of regulatory networks in hydrocarbon and lipid microorganisms. In: McGenity TJ, et al. (eds), Springer Protocols Handbooks. doi:10.1007/8623_2015_92
14. Chistoserdova L (2016) Systems biology tools for methylotrophs. Springer Protocols Handbooks. doi:10.1007/8623_2015_69
15. ETC Group (2007) Extreme genetic engineering: an introduction to synthetic biology. http://www.etcgroup.org/content/extreme-genetic-engineering-introduction-synthetic-biology
16. de Lorenzo V (2010) Synthetic biology: something old, something new. Bioessays 32:267–270
17. de Lorenzo V, Danchin A (2008) Synthetic biology: discovering new worlds and new words. EMBO Rep 9:822–827
18. Nørholm. Uracil-excision for assembly of complex pathways.
19. Porcar M, Danchin A, de Lorenzo V, Dos Santos VA, Krasnogor N, Rasmussen S, Moya A (2011) The ten grand challenges of synthetic life. Syst Synth Biol 5:1–9
20. Kelly JR, Rubin AJ, Davis JH, Ajo-Franklin CM, Cumbers J, Czar MJ, de Mora K, Glieberman AL, Monie DD, Endy D (2009) Measuring the activity of BioBrick promoters using an in vivo reference standard. J Biol Eng 3:4

21. Freemont P (2016) Protocol for the standardisation of transcriptional measurements. Springer Protocols Handbooks. doi:10.1007/8623_2015_148
22. de Las HA, Carreno CA, Martinez-Garcia E, de Lorenzo V (2010) Engineering input/output nodes in prokaryotic regulatory circuits. FEMS Microbiol Rev 34:842–865
23. de Lorenzo V (2011) Beware of metaphors: chasses and orthogonality in synthetic biology. Bioeng Bugs 2:3–7
24. Martinez-Garcia E, Benedetti I, Hueso A, De Lorenzo V (2015) Mining environmental plasmids for synthetic biology parts and devices. Microbiol Spectrum 3:Plas-0033-2014
25. Jaric J, Budisa N (2016) Design of orthogonal pairs for protein translation: selection systems for genetically encoding noncanonical amino acids in *E. coli*. Springer Protocols Handbooks. doi:10.1007/8623_2015_105
26. Cases I, de Lorenzo V (2005) Genetically modified organisms for the environment: stories of success and failure and what we have learned from them. Int Microbiol 8:213–222
27. Schmidt M (2016) Improving biocontainment with synthetic biology: beyond physical containment. Springer Protocols Handbooks. doi:10.1007/8623_2015_90

Protocol for the Standardisation of Transcriptional Measurements

Christopher D. Hirst, Catherine Ainsworth, Geoff Baldwin, Richard I. Kitney, and Paul S. Freemont

Abstract

A key component of the engineering approach underlying synthetic biology is the use of standardisation to enable better design of biological systems. One of the most important areas to standardise is the measurement of part, device and system activity in order to improve designs and aid sharing of data. While methods for standardising transcriptional measurements have been designed, they have suffered from poor uptake, and as more parts and systems are detailed, potentially useful information and comparison may be being lost. This protocol takes the best of the previously developed standards while adding some advice for best practice and data standardisation, designed to improve the ease with which data collected in separate labs may be shared and used. Standardisation of measurements and data has the potential to allow greater understanding of the biological systems synthetic biologists engineer and in turn lead to better tools to allow the design of larger and more complicated systems.

Keywords: Fluorescence, Standardization, Synthetic biology, Transcription, Transcriptional measurement

1 Introduction

Synthetic biology aims to utilise engineering approaches to aid the development of biological systems that function as initially designed. Of the many engineering principles that may be applied to advance these goals, one of the most important is standardisation both in terms of the definitions of physical pieces or DNA 'parts' and the procedures used to assemble and measure these parts. Significant work has been carried out to improve the standardisation of physical parts by descriptions and decoupling of their junctions [1, 2], the boundaries of the parts where they contextually interact with each other, and also in the area of improving and standardising the methods which can be employed to put these parts together [3–5]. When it comes to measurement standards however, progress has continued to be slow, while the number of

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 9–26,
DOI 10.1007/8623_2015_148, © Springer-Verlag Berlin Heidelberg 2015, Published online: 18 November 2015

projects and characterisations of parts and in particular libraries of parts has been increasing rapidly [2, 6, 7]. While these projects have produced much useful data, it is difficult at best to compare these datasets in meaningful ways and enable the production of the part and device models required for the field to achieve its aims.

One of the most important parts required for system design are the promoters which trigger transcription as most biological systems require the expression of various proteins or RNAs. While the immediate product of these parts is RNA molecules, the RNAs themselves can be difficult to study directly as this often requires lysis of the host and purification of mRNAs for use by methods such as RT-PCR [8]. This has meant that the measurement of promoter parts has tended to be carried out by monitoring production of fluorescent proteins encoded on those mRNAs. This has allowed characterisation experiments to be carried out in large numbers through the use of microplates and plate readers but at a cost of no longer directly assaying the signal of interest.

While protein-based experiments have been sufficient for now, many new tools and techniques have recently been developed to improve characterisation, each with advantages and disadvantages. New sensors based on RNA capable of directly reporting mRNA levels have been demonstrated for characterisation purposes [9]. These sensors have been observed to produce somewhat weak signals (necessitating flow cytometry or microscopy measurement) [9], but this may be a difficulty alleviated through improvements in aptamer chemistry [10]. Developments with in vitro transcription-translation technology have improved the reliability of these systems [11] to the point where it has been demonstrated that data collected in vitro may be comparable to in vivo derived data [12]. These systems are much less sensitive to context but also have short life spans and may struggle to execute complicated designs or systems. Finally, there has been increased use of flow cytometry either alone [13] or in combination with other techniques such as RNA-seq [7]. Flow cytometry provides high-quality expression data for individual cells but generally at a cost of throughput because of the time to assay sufficient cells for analysis.

This proliferation in technologies could be a problem for standardisation of transcriptional measurement; however, there are many areas where these different methodologies can be standardised to produce data which is as reliable and comparable as possible. For simplicity this chapter will focus primarily on the methods for standardising in vivo protein-based measurements as these are the most common, and the methods explained can be easily adapted to in vitro and aptamer-based methodologies. These standardisation methods focus on converting signal observed into standard units such as those described by Canton et al. [14] and Beal et al. [13] or by use of a reference system [15, 16]. As these techniques work in different ways, they are somewhat

complementary, and it may be advised to use both to gather insight into the measurement. In addition standardisation of the measurement scenario will also be discussed as this will significantly improve the reliability and reproducibility of such experiments.

2 Protocol

2.1 Approach Part 1: Measurement Environment and Reference Standards

If the environment where a final design may operate is known, it would be beneficial to replicate this during the measurement process. Where this is not the case of transcriptional measurement being carried for a more generic scenario or to obtain data for a set of parts or devices, a more standard measurement environment may be more suitable.

Biological systems operate in a chassis upon which they are dependant. As the properties of the chassis will alter the observed measurement, standardisation of the measurement environment should be considered in order to aid comparison and reuse of measurement data. When transcription within a chassis is measured, it requires the use of the host's machinery to produce the transcript and any reporter. This machinery minimally includes the host's RNA polymerase and ribonucleotide resources but often also ribosomal machinery when a protein reporter is used. The amount of these pieces of machinery will differ between hosts with evidence from in vivo studies that changing strain of organism alone significantly alters results [17]. Within a living host, the growth rate can change in response to a number of environmental factors and has been observed to alter a number of global properties inside host cells. This not only includes the relative abundance of RNA and protein molecules, the rate of protein synthesis and transcription which is being measured [18] but also the copy number on which many plasmids are kept [19]. Although in vitro transcription-translation mixes do not possess a growth rate which could alter the abundance of the machinery, the source of the mix currently is a living chassis, and as a result the process of generating the mix could have similar impacts.

As such the standardisation of any transcriptional measurement must consider both which hosts are appropriate for testing and in what environment (media and vessel) in which the measurement should take place. Many labs appear to use one of many cloning strains for the measurement of transcription, and the reasons for this choice are unclear but may be simply because it is easier than transferring to a new strain for the measurement. Additionally the requirement by some projects to use knockout strains further makes the restriction to a single standard strain of host for measurement impractical. Some strains are more appropriate candidates for standardised measurements, a few of which have been frequently used for the transcriptional measurements documented to

Table 1
Suitable strains and media for standardised transcriptional measurements

Media or strain	Notes and suggested reasons for use
Suggested strains	
MG1655	*E. coli* lab strain frequently used for its relatively 'wild' genotype. Frequently used for transcriptional measurements
MDS42	Most minimalised genome *E. coli* currently available that still exhibits relatively 'normal' behaviour
BW25113	Source strain for the Keio collection so useful when knockouts are required. Frequently used for measurements
DH5∝	Cloning strain commonly used in transcriptional measurement
BL21	T7 RNA polymerase-carrying strain often used for protein expression. Frequently used for measurements and often used for the generation of TX-TL in vitro mixes
Media	
Rich MOPS EZ media	Rich variant MOPS *E. coli* media, commercially available and modular in nature to allow for easy replacement of components
Minimal MOPS media	Minimal version of MOPS media, based on only the MOPS mixture and potassium thiamine. Also commercially available

date (Table 1), and as some of the absolute unit calibration techniques take into account the cell strain, it would be advisable to use the most appropriate suggested strain if a choice of strain is possible.

The media used however is much more open to standardisation, and experience from other biological fields should be a strong guide towards the selection of highly or better yet entirely defined media. While many commonly used media components are given the same name or are made by the same method, there may be subtle differences that could influence a measurement. For this reason, fully defined or standardised media should be used if possible, and at present the only fully defined, widely used media is the EZ Rich [20] or minimal [21] variants of the MOPS media first developed by Neidhardt [22]. Any media where all the components are at known concentration would also be appropriate but these formulations need to be well documented. Where the use of a fully defined media is inappropriate (e.g. in a bioreactor), the use of particularly variable media such as LB or those based on variable components such as tryptone or yeast extract should be avoided. A suggested set of standard *E. coli* strains and media can be seen in Table 1, and similar standards would ideally be established for other organisms when they are regularly used by the community. Where in vitro mixes are made, it may be suitable to generate these from one of the suggested strains where possible.

For in vivo measurements, beyond the strain and the media, other factors can also influence measurements. Biological systems are inherently in a constant state of flux, and for this reason it

cannot be assumed that they continually stay the same, particularly during growth. For this reason, care should be taken to ensure that samples are always measured in the same state, generally governed by the number of cells, and ideally multiple measurements should be taken during growth so as to indicate any variation in results and to yield data more likely to be consistent. This however may not be possible depending on equipment available. With methods such as flow cytometry and RNA-seq where obtaining a time series of data is difficult, expensive or impractical, it is therefore important to keep growth consistent among samples. The strategy employed to obtain this consistency will depend upon the strains and media used but will likely require volumetric of concentration-based dilution and set-up steps.

A complementary standardisation technique can be employed where a standardised measurement set-up cannot be used. Arguably it would ideally also be used with every measurement to indicate any abnormalities or unknown deviations in the measurement procedure or conditions. This strategy is the measurement of a reference construct which is maintained across experiments and when carefully designed allows the derivation of transcriptional output in a standardised unit. This method is based on the notion that while conditions may affect how a part functions, its behaviour in relation to other similar parts should remain the same; this can be used up to define the behaviour of a part relative to other parts of the same class and was the core of the relative promoter unit (RPU) standard [15]. The relative promoter unit standard used a near identical reference plasmid to dictate a value of 1 on an otherwise arbitrary scale around which all other measurements could be understood. Variants on this procedure have since been demonstrated in similar situations resulting in αRPU [23] and relative expression unit [24]-based results and have also been shown to be appropriate for in vitro experiments [16]. With careful design, these reference measurements can be useful for removing a large amount of context sensitivity but do not provide an absolute output which may be significantly more useful for models and design software.

To standardise a transcriptional measurement, this way a suitable reference construct must be measured as part of the assay. As an additional benefit, it is possible to compare the results of the reference construct between assays, which allows it to indicate where abnormalities in the measurement procedure may have occurred. This is particularly useful for in vitro assays where the transcription-translation mix can vary significantly batch to batch [16]. Generally, the reference construct is carried as a separate sample in the assay, but it has been suggested that a second fluorophore could be used to carry out the reference measurement. The reference constructs are suitable for both in vivo and in vitro work and may even provide a method to allow data comparison between in vivo and in vitro experiments [12]. For reference measurements

to convert results into standard units, the reference construct must be carefully designed.

The original RPU standard [15] only called for the reference construct to match the tested construct in the region from the 5′ end of the transcript to an undefined distance into the protein being used as a reporter. While this region is likely to have a large impact upon the measurement, there may be many other effects from both the testing construct and the overall plasmid or genomic environment upon which it is hosted. For this reason a very strict reference construct should be used. In this scenario, the reference should be identical with the only exception of the part being measured. If parts within the design (particularly a promoter where the transcription start site is within its sequence) are likely to be switched in future experiments it would also be advisable to use promoter/5'UTR decoupling tools [1, 25] and insulator sequences such as those for promoters [23]. An SBOL diagram of such a construct is included in Fig. 1. The mathematical method to produce results in relative units is included with mathematical methods for the other output units in the data analysis section.

2.2 Approach Part 2: Equipment Calibration

Most measurements of transcriptional activity are based around the use of fluorescent proteins and so are carried out on only a few of pieces of equipment. As a result, standardising around these pieces of equipment is a viable option. The equipment used in many of the transcriptional measurements are plate readers and flow cytometers. Both types of equipment gather fluorescence data about samples that are given in arbitrary fluorescence output units and optical density or absorbance values which are also effectively arbitrary. It is however possible to convert these outputs into results with standard units (Table 2). Results converted this way can be compared instantly. In addition these methods can allow the calculation of more complicated results types, such as polymerases per second (PoPs) [14].

2.2.1 Plate Reader Fluorescence Experiments

For in vivo transcriptional assays, plate readers monitor the emission of fluorescent light to judge the amount of fluorescence emitted by protein within cells as well as the size of the bacterial population responsible for this signal based on the amount of light they scatter or absorb. As an in vitro reaction uses a fixed volume of reaction mix, such transcriptional assays only require the emission of fluorescence to be monitored. Both scales are inherently arbitrary and will vary from machine to machine but can be converted into absolute units.

The absolute unit for fluorescence signals are molecules of fluorescent protein. This calibration was originally designed by Canton et al. [14] and will yield results in terms of the molecules actually produced by the transcription and translation processes, and so these results may be more immediately useful in models or

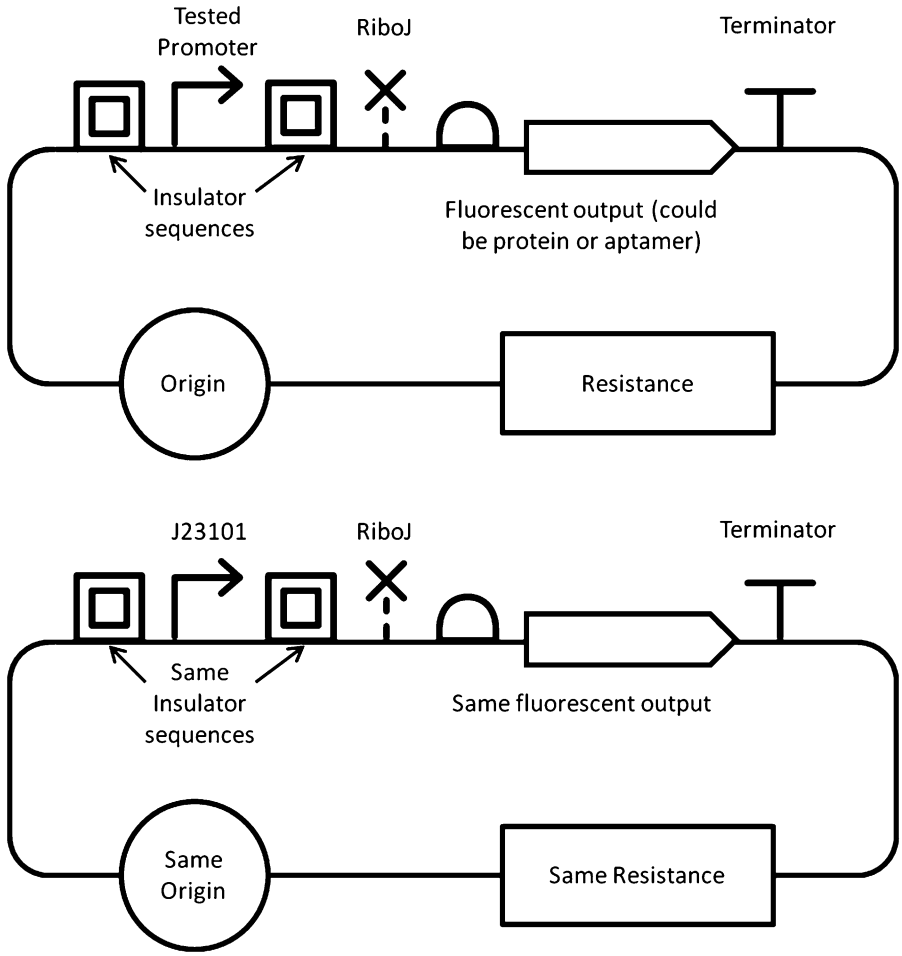


Fig. 1 Example of a suitable reference control and integration of promoter part standards. As many biological 'parts' may have unknown interactions with each other or their hosts for accurate comparison, the reference construct should be as close to identical to the tested construct as possible within the experimental constraints. An example constitute promoter testing construct and its ideal reference are shown in SBOL-v notation [26]. Surrounding the promoter in the above diagram are two part standardisation devices, two promoter insulators [23] and an mRNA cleavage device [1]. Use of such devices may be useful for measurement accuracy (depending upon experimental context)

design tools. It should be possible to use these methods to perform a calibration for molecules of fluorescent aptamer or similar molecules. For increased reliability, it is advisable to regularly run a test plate or chemical standard on the plate reader to ensure there that response of the detector does not change over time or as a result of hardware use. This is important as the equipment calibrations are likely to be only carried out rarely (particularly molecules of fluorescent protein calibrations). Additionally, it may be sensible to carry out these calibrations for a small number of different gains or sensitivities so that the calibration procedure does not need to be

Table 2
Possible and appropriate units for standardised transcriptional measurements using commonly used equipment

Measurement equipment	Output units	Experimental requirement
Plate reader (fluorescence)	Arbitrary units[a] Molecules of fluorescent protein	None Calibration curves for Fluorescence from purified protein (in lysate) Cell physiology on fluorescent protein
Plate reader (absorbance)	Absorbance/optical density[a] Colony-forming units Cell population numbers	Linearity calibrations Plate counting assays Cell bead counting calibration
Flow cytometer (fluorescence)	Arbitrary units[a] Equivalent units of fluorescein/ fluorescent dye	None Calibration bead data

[a]Denote result unit which is not standardised and so cannot be easily compared without referencing

repeated and one of these pretested gains used instead. If the change in detector properties is carefully monitored, it should be possible to adjust calibrations to take this into account. It is important to note that the fluorescence detection part of these calibrations must be carried out at the same temperature as that used for the transcriptional measurement due to the affect of temperature upon fluorescent signals.

The conversion of fluorescence results into molecules of fluorescent protein is slightly different for in vivo and in vitro measurements (the alteration will be detailed following the in vivo protocol). For conversion of in vivo data, two calibrations are required, the fluorescence obtained from known concentrations of protein in cell lysate (produced chemically) and the change in fluorescence observed from fluorescent protein when the cells are chemically lysed. The second conversion is required because the exact nature of the environment inside the tested cells is unknown, with some properties such as pH [27] able to affect the signal from a fluorescent protein. The only way to deal with this unknown effect and ensure accuracy of results is to use a common environment (chemically generated cell lysate). For both of these calibrations, controlled cell lysis is critical, and this should be achieved by chemical methods as this will ensure >95% cell lysis. The following has been demonstrated using B-PER II lysis buffer (Pierce), and if other lysis solutions are to be used, they should be tested for lysis efficiency by plating (see Sect. 2.2.2). As the environment inside

cell strains may be different if a new strain is used, the calibration should be repeated.

For generation of lysate, cells should be grown to an approximate OD of 0.5 prior to spinning at 3,000 × *g* for 5 min. Cell solutions need to be kept on ice following this step of the process. Following spinning cells should be resuspended in PBS to an OD of 0.5 before being spun and resuspended in PBS a second time. Protease inhibitors should be added to the cells in PBS and then an equal volume of lysis solution added to generate cell lysate (this should be done in tubes if possible to reduce the possibility of generating bubbles in wells). For the calibration of the change in fluorescence signal in lysate versus intact cells (i.e. accounting for the cellular environment), batches of cells expressing differing levels of fluorescent protein should be spun individually, washed with PBS and then stored on ice prior to generation of lysate. The individual samples should be equally split in two and one of each pair lysed using an equal volume of lysis buffer, while the other is diluted with an equal volume of PBS. Samples of each should then be transferred to a microplate (being careful not to introduce bubbles) and read for fluorescence. Samples not expressing fluorescent protein should be included to account for autofluorescence. The signal from these autofluorescence controls should be subtracted, and the relationship between the fluorescence observed before and after lysis should be calculated for use in the overall calibration.

The protein molecule calibration is itself also a two-step procedure (shown in Fig. 2), first requiring purification and quantification of protein and secondly detection of fluorescent signals from known concentration samples. The purification should be carried out by commonly used protein purification set-ups which allow purification while minimally affecting the protein such as his-tagging. Cleavage of a tag may be beneficial but could also cause problems related to folding of resulting protein or sample contamination. Protein purity should ideally be >90%; however, as long as the purity can be ascertained, that should be acceptable. For quantification a small dilution series should be produced by diluting protein in PBS to improve accuracy of quantification and to identify any pipetting problems arising due to viscosity of protein sample. Quantification of the protein should then be achieved by using a commercially available protein quantification kit to quantify the dilution series of purified protein rather than nanodrop or similar methods as the fluorescent properties of the proteins may interfere with these readings. A large range of dilutions may then need to be created from these samples by careful pipetting if they do not cover a large enough fluorescence range. To generate the final calibration curve, 40 μl of these samples should be added first into microplate wells before addition of 160 μl of nonfluorescent cell lysate generated by the above methods. Samples containing no fluorescent

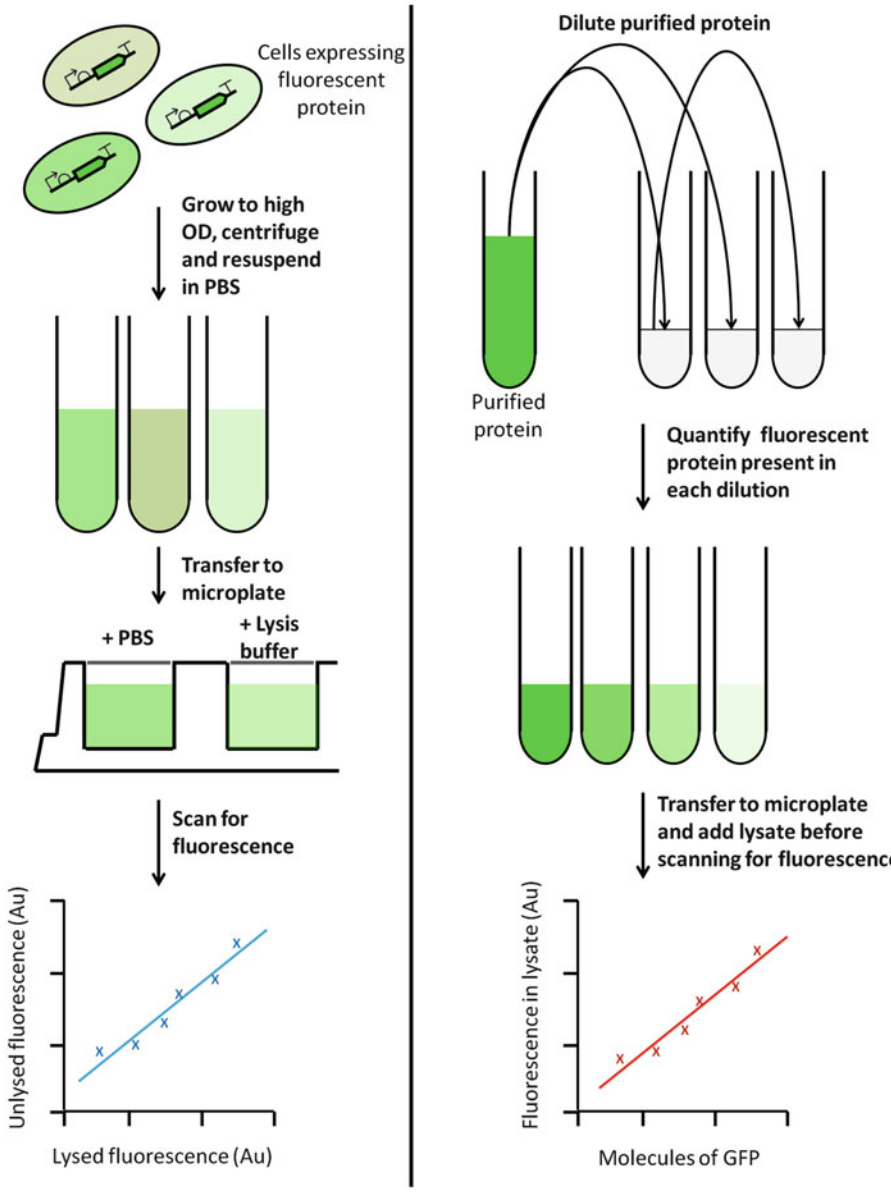


Fig. 2 Method to obtain calibration of observed fluorescence into molecules of fluorescent protein. The experimental workflows shown produce two separate calibrations relating either the fluorescence signal inside cells or the number of molecules of fluorescent protein in a sample to the fluorescence observed in cell lysate. By carrying out these calibrations with protein in cells and purified proteins separately, it is possible to convert via the common scenario (protein in cell lysate) from the fluorescence observed by a plate reader to the number of molecules of protein present inside cells in a microplate well

protein should also be added to use as a background measurement. Following fluorescent measurement of the GFP in lysate, the fluorescence background from the lysate-only samples should be subtracted, and the number for molecules of protein added to each well should be calculated. From these results, the mathematical

relationship between molecules of fluorescent protein and observed fluorescence should be calculated and can then be used to convert fluorescent results into molecules of fluorescent protein (*see* Sect. 3 for details).

For in vitro measurements, only the signal observed from quantified protein is required although this should be carried out in in vitro reaction mix to maintain the chemical properties that will be observed under measurement conditions. For fluorescent reporters that are not proteins (such as fluorescent RNAs), a similar method using the reporter molecule may be appropriate but would need thorough testing.

2.2.2 Cell Population Measurement Calibration

Collection of in vivo transcriptional measurements on a plate reader also requires the determination of the size of the cellular population. This is normally established by measuring the scattering of a beam of light. While many pieces of equipment measure this scattering and produce absorbance or optical density results, the result observed for different pieces of equipment varies widely. The absorbance of a liquid on one machine will not be the same as the absorbance measured on another, and this is made significantly worse by equipment 'linearity' where the measurement equipment no longer detects a doubling in the concentration of the absorbing material as a doubling in fluorescence. For this reason, the experimental equipment used to take these measurements should also be calibrated for their absorbance results. It would be advisable to keep all arbitrary measurements in absorbance so as to avoid confusion with optical density and pathlength corrections. While a dye exhibiting known optical densities at various concentrations could be used, it is almost as simple and potentially more useful to calibrate these measurements directly into the number of cells in a given sample vessel. This calibration should be carried out by taking samples from growing cultures at regular intervals and assaying for the number of cells by one of two different methods depending on the equipment available. While both of these methods are viable, if the lab has access to flow cytometry equipment, a method employing counting beads should be used as this will give more accurate data than plating assays.

To ensure accurate calibration of cell numbers, the experiment must be set up using conditions identical to those under which the measurement was (or will be) taken. In particular the cell strain, growth medium, growth vessel (including volume) and incubation settings should be kept the same unless an experiment demonstrates there is no difference in cell size as a result of these factors. Additionally, the plasmid held by the host should also carry the same resistance and cause a similar burden (if applicable) to the measured system. For bead counting assays, the inclusion of a strong fluorescent marker on the sample bacteria will make gating

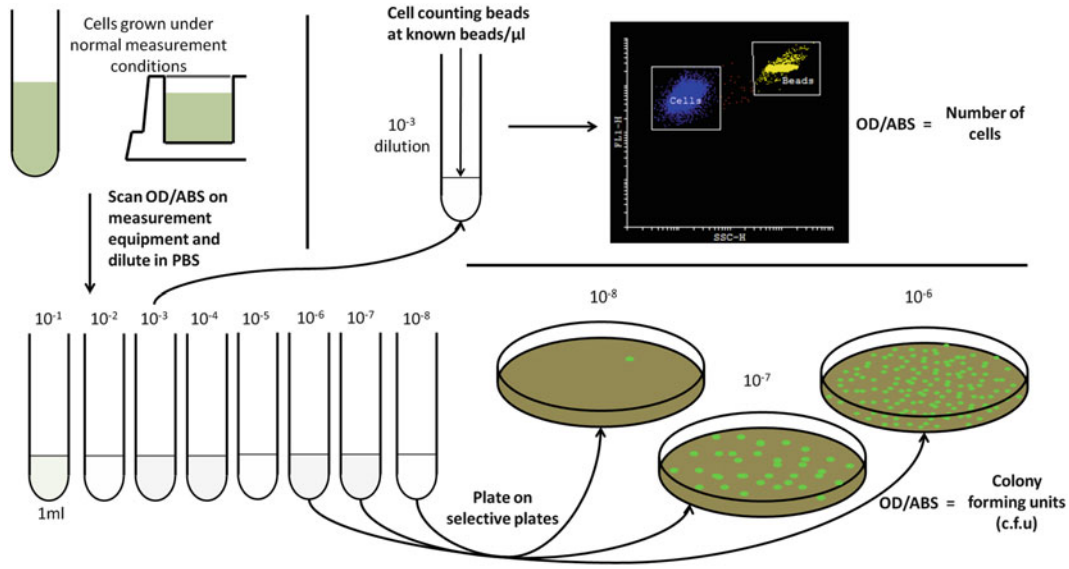


Fig. 3 Methods for determining cell numbers from absorbance or optical density data. Diagrams for the workflow of plate or bead counting are displayed, highlighting the common sample preparation and dilution steps before differing dilutions can be taken for either type of cell number counting. Low dilutions can be taken for bead counting, where samples are run through a flow cytometer with beads and the number of cells calculated using the ratio of cells to beads and the dilution run through the flow cytometer. Larger dilutions are required for plating assays, where following spreading on agar plates, the number of colonies on the most suitable plate is counted and multiplied by the dilution factor to obtain the number of colony-forming units (c.f.u.) in the original sample

significantly simpler. From this set-up, samples should be taken and measured for absorbance/optical density before being diluted to a level appropriate for either plating or bead counting. Multiple samples should be taken to observe the relationship between absorbance and cell numbers or colony-forming units. For plating assay, this should be done multiple times to account for the possible errors as a result of the large number of dilutions. For a diagram of the two methods, *see* Fig. 3.

For plating assays, large dilutions should be carried out by serial dilution using well-calibrated pipettes to yield dilutions in the range 10^6–10^8. These dilutions should then be plated on selective plates and grown overnight. The following day the plates should be inspected for colonies, and the colonies on the lowest dilution plate that can easily be counted should be counted. This number of colonies should then be multiplied by the dilution (being careful to remember any dilution which may have occurred during plating) to obtain the colony-forming units of the original sample. The colony-forming units are the number of viable cells in the sample and so should approximate the population size in properly designed experiments.

For bead counting assays, smaller dilutions should be carried out. The dilutions here should be suitable to yield cell numbers to be assayed that are ideally within an order of magnitude of the number of beads in the counting mixture. As such the exact dilution will depend on the sample absorbance and the concentration of counting beads being used, though it is likely to be in the region of 40- to 1,000-fold (for absorbance values in the range of 0.002–1 and a bead concentration of approximately 6,000/μl). It should be possible to establish this with the first sample (or a test sample of known absorbance) and from then on use educated guesses based on the observed absorbance and number of event measured for the previous sample. For the counting itself, as many events as are sensibly possible should be recorded per sample to obtain the most accurate results. Flow cytometry results should be gated (ideally in side scatter and fluorescence) to distinguish bead events, cell events and other events. The ratio of the bead and cell populations should be calculated before multiplying by the number of beads per microlitre (to obtain cells/μl) and the dilution to obtain the number of cells in the original sample.

When the c.f.u. or counted cell numbers are calculated for multiple absorbance/optical density values, the absorbance results should be plotted against c.f.u. to obtain a mathematical relationship which can be used for calibration of all results into c.f.u. or cell number. This equation should be used to convert all absorbance or optical density results into c.f.u. or cell number following blanking of measurement (*see* Sect. 3 for more details) (Fig. 3).

2.2.3 Generation of Results in Polymerases per Second (PoPs)

Canton et al. [14] demonstrated that it was possible to convert results in absolute protein units into PoPs. While these units may be an ideal metrology for a transcriptional process, the full methods required to allow this conversion are clear. In short, several mRNA and protein parameters are required in order to calculate the PoPs output of the promoter using an ODE model based on transcription and translation. The parameters required are the mRNA degradation rate, the protein synthesis rate (from mRNA), the protein maturation rate and the degradation rate of immature protein (the full protein degradation rate would be required if the protein degrades after folding). The methodology required to obtain these parameters is not so clear, and anyone wanting to produce results in PoPs would be advised to look at the work of Canton et al. [14] for more details.

2.2.4 Flow Cytometry Measurement Calibration

Flow cytometry has recently begun to be used regularly for transcriptional measurements in a synthetic biology context. A large set of tools has been developed to work with this methodology, and part of this is an easily applicable method of unit standardisation [13]. Commercially available flow cytometry calibration beads are

available from many suppliers. Standardisation with these beads is very simple but is the only unit conversion which must be performed each time data is collected (whenever the cytometer is turned off then restarted). Prior to data acquisition, calibration beads must be run through the cytometer and data collected with all the fluorescence instrument settings to be used for data collection. If more than one fluorophore is to be used in a sample, care also must be taken to remove any signal caused by bleed from one channel into another (e.g. green fluorescence appearing in a red channel).

Many manufacturers offer calibration beads which possess multiple standard levels of appropriate fluorescent dyes (e.g. example fluorescein for GFP). The exact procedure varies but generally should begin with generation of a dye calibration curve. For the bead sample, non-bead event should be filtered out by gating, and then a manufacturer-provided calibration data (i.e. amount of dye per bead) should be applied to yield a conversion equation for arbitrary fluorescence signal to molecules of equivalent fluorophore (e.g. MEFL – molecules of equivalent fluorescein for GFP). This may have to be carried out without or after background signal removal as dictated by the bead manufacturer. Transcription measurement data should then be treated the same way as the beads with regard to background signal before the arbitrary fluorescence values observed are converted to MEFL or similar units via the bead calibration curve.

3 Data Analysis

Careful data analysis can be the difference between poor quality, noisy data and highly accurate insightful results, particularly when applied to the plate reader-based population measurements that are commonly reported. Care must always be taken to appropriately remove background signal prior to calibration of data and calculation of results. Additionally, great care should be taken with the determination of error and when to carry out data averaging as performing this at the wrong stage will artificially increase the observed error. While plate reader and flow cytometry data will be separated here, some data from the plate reader may be required for calculation of flow cytometry results (notably population growth rate).

Following the data analysis, care should be taken to be clear which format and unit the results have been produced in to allow accurate comparisons to be drawn between other datasets. If the data is to be put into a shared repository or sent to other groups, it would be highly beneficial to send the data along with information relating to context and analysis. At the bare minimum, this should include the DNA sequences used in the measurement, the assay media and the cell strain but could ideally be assembled to in some

kind of datasheet, with good previous examples including the F2620 datasheet [14] and BglBrick vector datasheets [28].

3.1 Plate Reader Data

This section will focus primarily on in vivo data as this is more complicated. For in vitro data, the main steps are the same except there are no absorbance measurements and autofluorescence should be handled by timepoint by timepoint subtraction of negative control fluorescence values. For in vivo plate reader data, data analysis should always begin with the absorbance measurements as they will require adjustment in order to estimate autofluorescence. First media absorbance values should be checked to ensure there has been no growth consistent with contamination. Small gains in absorbance values for the media may occur, but significantly they should ultimately plateau and must not increase exponentially. The average absorbance value of all the media wells which pass this examination should then be subtracted on a timepoint by timepoint basis from all sample absorbance to obtain corrected absorbance values.

The autofluorescence is dependent upon media and is also related to the population of cells (as measured by absorbance) in many media. It may be beneficial to carry out the conversion of absorbance signal into cell number of c.f.u. prior to calculating autofluorescence. To calculate the autofluorescence, the fluorescence and the corrected absorbance values for the negative controls are required. Autofluorescence should not be removed by removing the fluorescence of negative control cells at a given timepoint unless the absorbance values are very similar at that timepoint. Instead the relationship between the absorbance and fluorescence of negative cells should be identified by regression or similar mathematical analysis (different media tend to yield differing equation types but generally linear plus offset or quadratic are most appropriate). The corrected fluorescence can then be calculated by subtraction of the level of fluorescence expected to be observed from negative cells with the same absorbance.

For all output formats, both fluorescence and absorbance data should now be converted into a standard unit format to ensure that the results determined following this step are easily compared and should easily be reproduced. While for RPU calculations this may not be necessary as later steps will have the effect of cancelling out any conversions used, it is likely that having data in another absolute format will be beneficial. The conversions should be carried out by multiplying the data with the relevant conversion/calibration factors determined experimentally (*see* Sect. 2.2). With the data calibrated into standardised units, the results can be calculated in the most useful format for the desired objective, generally RPU or equivalent for relative results and synthesis rate or expression level (fluorescent molecule/cell). From the calibrated data, expression level or synthesis rate should be calculated first. Expression level is

the simplest to calculate but may require more analysis if a rate of transcription is required. These calculations are already carried out with analysis in synthesis rate, but results expressed this way will be noisier, particularly with weak transcriptional signals. If only a limited number of data points have been generated, the results should be presented in expression level, as synthesis rate ideally should be calculated from a few points to reduce some of the impact of noise and because it has been observed to change over time [15].

The expression level is the fluorescence signal normalised by the number of cells and should be carried out for each replicate of a sample individually prior averaging. If the expression level is the desired output, the resulting expression level for all replicates can then be averaged and the standard deviation calculated. Synthesis rate requires the calculation of the change in fluorescence normalised by the cell population. While many equations have been suggested for this calculation, the following equation taken from Kelly et al. [15] should be used:

$$\text{Synthesis Rate (Plate Reader)} = \frac{(\text{cFl}_i - \text{cFl}_{i-1})}{(\text{cOD}_i - \text{cOD}_{i-1})/2} / t_i - t_{i-1}$$

where t is time, cFl is calibrated fluorescence and cOD is calibrated optical density/absorbance. This should give results in the format such as molecules of protein per cell per minute or similar. As with expression level, this should be calculated for each individual replicate prior to averaging and calculation of standard deviation. As with expression level, the averaging of all replicates should only be carried out at this stage if synthesis rate is the desired output format to prevent unnecessary addition of error.

Output in relative promoter units and relative expression level can now be calculated from the synthesis rate and expression level, respectively. While the calculation to do this is simple (division of the sample expression level or synthesis rate by that of the reference construct), the ordering of division and averaging steps is important to avoid introduction of unnecessary error. As a first step, the replicates for both samples and reference constructs should be grouped according to their data run (i.e. the data for all the replicates obtained on day 1 of a 3-day experiment should be kept separate from those of the other days). At this point the reference results only should be averaged for each day. Following this the division of sample results by the average reference result of that data run should be carried out, yielding the relative expression level or RPU for all the sample replicates. These results should now be averaged and the standard deviation calculated.

3.2 Flow Cytometry Data

The flow cytometry data is much simpler to analyse following gating and many more accurate statistics can be produced. Owing to the difficulties associated with the one box on a log scale,

software that can handle data with a 'logicle' axis should be used for gating and determination of flow cytometry statistics [29]. Flow cytometry should always be gated in forward and side scatter as tightly as possible before calculation of results. Dimensions of this gate will depend upon instrument settings, host species and strain and growth phase. Generally following gating, the geomean is the key statistic to be determined for each individual replicate, but on some occasions the mean or median statistic may be more appropriate (usually as the result of subpopulations or significant background noise). It may be advisable to also calculate the coefficient of variation as this is a better estimation of the variation in results (the geomean/mean/median will be used primarily to calculate error and only the geomean will be referred to from this point).

As flow cytometry measures the fluorescence per cell, to calculate the expression level, the only requirement is to remove the background signal. This can be carried out by subtracting the average geomean of the negative controls from the geomean of each sample. Depending on the supplier of calibration beads, results should be converted into molecules of equivalent dye either immediately before or after the background subtraction step. If the synthesis rate is desired, it can be calculated from this calibrated data by division of the doubling time (usually determined from plate reader or optical density data) to give results in units of molecules of equivalent dye per cell per minute. Alternatively, the calibrated result is a standardised expression of the fluorescence per cell. The relative expression level and output in relative promoter units can be calculated using the same equations as the plate reader data (if the OD terms are ignored) following background signal removal.

Acknowledgements

We would like to acknowledge Jake Beal for useful discussions regarding the standardisation of flow cytometry results. We also thank EPSRC for funding and colleagues in CSynBI particularly Guy Bart-Stan and Tom Ells.

References

1. Lou C, Stanton B, Chen Y-J, Munsky B, Voigt CA (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. Nat Biotechnol 30(11):1137–1142
2. Mutalik VK, Guimaraes JC, Cambray G, Lam C, Christoffersen MJ, Mai QA et al (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. Nat Methods 10(4):354–360
3. Casini A, Macdonald JT, Jonghe JD, Christodoulou G, Freemont PS, Baldwin GS et al (2014) One-pot DNA construction for synthetic biology: the Modular Overlap-Directed Assembly with Linkers (MODAL) strategy. Nucleic Acids Res 42(1), e7, Epub 2013/10/25
4. Werner S, Engler C, Weber E, Gruetzner R, Marillonnet S (2012) Fast track assembly of multigene constructs using Golden Gate cloning and the MoClo system. Bioeng Bugs 3 (1):38–43, Epub 2011/12/01
5. Torella JP, Lienert F, Boehm CR, Chen JH, Way JC, Silver PA (2014) Unique nucleotide

sequence-guided assembly of repetitive DNA parts for synthetic biology applications. Nat Protoc 9(9):2075–2089, Epub 2014/08/08

6. Cambray G, Guimaraes JC, Mutalik VK, Lam C, Mai QA, Thimmaiah T et al (2013) Measurement and modeling of intrinsic transcription terminators. Nucleic Acids Res 41 (9):5139–5148
7. Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP et al (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. Proc Natl Acad Sci U S A 110(34):14024–14029
8. Chappell J, Freemont P (2013) In vivo and in vitro characterization of sigma70 constitutive promoters by real-time PCR and fluorescent measurements. Methods Mol Biol 1073:61–74, Epub 2013/09/03
9. Pothoulakis G, Ceroni F, Reeve B, Ellis T (2014) The spinach RNA aptamer as a characterization tool for synthetic biology. ACS Synth Biol 3(3):182–187
10. Strack RL, Disney MD, Jaffrey SR (2013) A superfolding spinach2 reveals the dynamic nature of trinucleotide repeat-containing RNA. Nat Methods 10(12):1219–1224, Epub 2013/10/29
11. Sun ZZ, Yeung E, Hayes CA, Noireaux V, Murray RM (2014) Linear DNA for rapid prototyping of synthetic biological circuits in an Escherichia coli based TX-TL cell-free system. ACS Synth Biol 3(6):387–397
12. Chappell J, Jensen K, Freemont PS (2013) Validation of an entirely in vitro approach for rapid prototyping of DNA regulatory elements for synthetic biology. Nucleic Acids Res 41 (5):3471–3481
13. Beal J, Weiss R, Yaman F, Davidsohn N, Adler A (2012) A method for fast, high-precision characterization of synthetic biology devices. MIT-CSAIL-TR-2012-008. 2012(008). Epub 2012/4/7 http://hdl.handle.net/1721.1/69973
14. Canton B, Labno A, Endy D (2008) Refinement and standardization of synthetic biological parts and devices. Nat Biotechnol 26(7):787–793, Epub 2008/07/10
15. Kelly JR, Rubin AJ, Davis JH, Ajo-Franklin CM, Cumbers J, Czar MJ et al (2009) Measuring the activity of BioBrick promoters using an in vivo reference standard. J Biol Eng 3:4, Epub 2009/03/21
16. Takahashi MK, Chappell J, Hayes CA, Sun ZZ, Kim J, Singhal V et al (2015) Rapidly characterizing the fast dynamics of RNA genetic circuitry with cell-free transcription–translation (TX-TL) systems. ACS Synth Biol 15 (4):503–515
17. Cardinale S, Joachimiak MP, Arkin AP (2013) Effects of genetic variation on the E. coli host-circuit interface. Cell Rep 4(2):231–237, Epub 2013/07/23
18. Klumpp S, Zhang Z, Hwa T (2009) Growth rate-dependent global effects on gene expression in bacteria. Cell 139(7):1366–1375, Epub 2010/01/13
19. Klumpp S (2011) Growth-rate dependence reveals design principles of plasmid copy number control. PLoS One 6(5), e20403, Epub 2011/06/08
20. University of Wisconsin E. coli Genome Project. EZ Rich Defined Medium. 2002 [updated 5/5/2003; cited 2014 01 May]. http://www.genome.wisc.edu/resources/protocols/ezmedium.htm.
21. University of Wisconsin E. coli Genome Project. MOPS Minimal Medium. 2002 [updated 5/5/2003; cited 2014 01 May]. http://www.genome.wisc.edu/resources/protocols/mopsminimal.htm.
22. Neidhardt FC, Bloch PL, Smith DF (1974) Culture medium for enterobacteria. J Bacteriol 119(3):736–747
23. Davis JH, Rubin AJ, Sauer RT (2011) Design, construction and characterization of a set of insulated bacterial promoters. Nucleic Acids Res 39(3):1131–1141, Epub 2010/09/17
24. Temme K, Zhao DH, Voigt CA (2012) Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. Proc Natl Acad Sci U S A 109(18):7085–7090
25. Qi L, Haurwitz RE, Shao W, Doudna JA, Arkin AP (2012) RNA processing enables predictable programming of gene expression. Nat Biotechnol 30(10):1002–1006
26. Quinn J, Beal J, Bhatia S, Cai P, Chen J, Clancy K, et al. Synthetic Biology Open Language Visual (SBOL Visual), version 1.0. 0. 2013
27. Kneen M, Farinas J, Li YX, Verkman AS (1998) Green fluorescent protein as a noninvasive intracellular pH indicator. Biophys J 74 (3):1591–1599
28. Lee TS, Krupa RA, Zhang F, Hajimorad M, Holtz WJ, Prasad N et al (2011) BglBrick vectors and datasheets: A synthetic biology platform for gene expression. J Biol Eng 5:12, Epub 2011/09/22
29. Herzenberg LA, Tung J, Moore WA, Parks DR (2006) Interpreting flow cytometry data: a guide for the perplexed. Nat Immunol 7 (7):681–685, Epub 2006/06/21

Uracil Excision for Assembly of Complex Pathways

Ana Mafalda Cavaleiro, Morten T. Nielsen, Se Hyeuk Kim, Susanna Seppälä, and Morten H.H. Nørholm

Abstract

Despite decreasing prices on synthetic DNA constructs, higher-order assembly of PCR-generated DNA continues to be an important exercise in molecular and synthetic biology. Simplicity and robustness are attractive features met by the uracil excision DNA assembly method, which is one of the most inexpensive technologies available. Here, we describe four different protocols for uracil excision-based DNA editing: one for simple manipulations such as site-directed mutagenesis, one for plasmid-based multigene assembly in *Escherichia coli*, one for one-step assembly and integration of single or multiple genes into the genome, and a standardized assembly pipeline using benchmarked oligonucleotides for pathway assembly and multigene expression optimization.

Keywords: BioBricks, DNA editing, Metabolic engineering, Molecular cloning, Synthetic biology, Uracil excision cloning

1 Introduction

The polymerase chain reaction (PCR) [1] is a simple yet incredibly powerful technology that revolutionized molecular biology. Shortly after the advent of PCR, a handful of methods for assembly of PCR-amplified DNA into larger constructs was developed. PCR generates double-stranded DNA flanked by sequences that are defined by the two PCR primers, and several methods exist that facilitate the formation of cohesive ends for specific higher-order assemblies (Fig. 1). Simple features can be added when the oligonucleotides are chemically synthesized. As an example, uracil excision DNA assembly makes use of oligonucleotides where selected thymines are replaced by uracils. This is a non-mutagenic and PCR-tolerated replacement, as the uracil is able to form base pairs with adenine nucleotides on the complementary strand [2–4]. Following PCR, the uracils are selectively removed by treatment with uracil DNA glycosidase, leaving a chemically unstable phosphoribose backbone. At elevated temperatures, the upstream sequence

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 27–37,
DOI 10.1007/8623_2015_133, © Springer-Verlag Berlin Heidelberg 2015, Published online: 08 September 2015

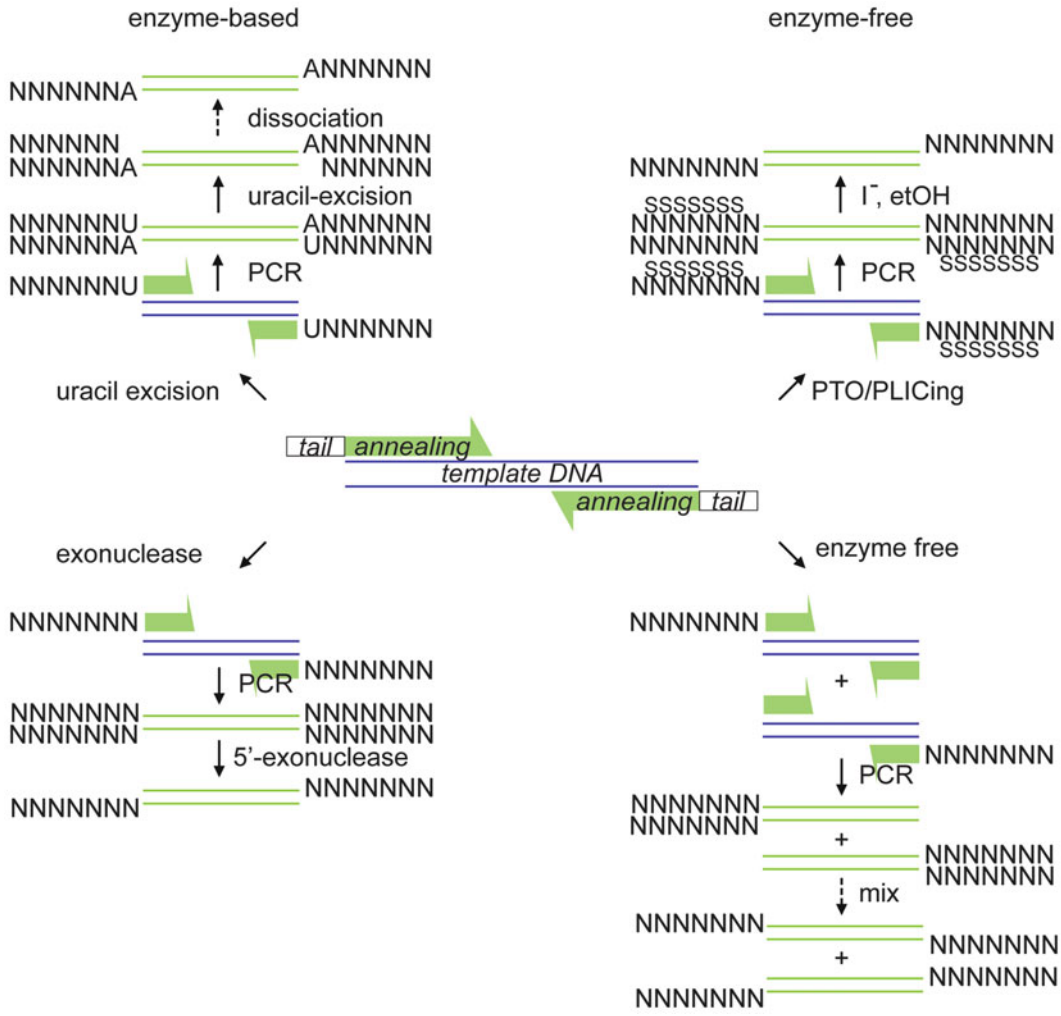


Fig. 1 Illustration of different methods to create cohesive ends on PCR fragments for specific higher-order assemblies. In the schematic examples, all four methods can generate the same 3′ cohesive ends that are not filled in by excess DNA polymerase activity from the PCR. Thus, all methods can in principle be employed directly after PCR with no prior purification. *S* denotes the phosphorothioate modification employed in the PTO/PLICing cloning technology. I^- denotes iodine and etOH denotes ethanol. For more information and references, see main text

dissociates, generating a single-stranded DNA overhang. A recently developed similar approach uses phosphorothioate (PTO)-modified synthetic oligonucleotides [5]. PTO-modified DNA is converted to single-stranded DNA by treatment with a solution of iodide and ethanol. Thus, in the case of PTO, the formation of cohesive ends is enzyme-free. Another enzyme-free route to cohesive ends on PCR products involves the use of two pairs of highly similar oligonucleotides, but of slightly different length, for amplification of the same DNA template [6]. When the resulting two

PCR products are mixed, denatured, and reannealed, the two products recombine and form single-stranded ends defined by the length difference of the two oligonucleotide pairs. However, this approach complicates the PCR setup (and doubles the price tag) and does not seem to be extensively used. Finally, exonuclease-catalyzed recessions of the ends of the DNA are heavily used alternatives, e.g., in the form of ligase-independent cloning (LIC) [7] or the commercially available cloning kit Gibson Assembly [8].

In our experience, uracil excision excels in robustness, simplicity, and price tag. This may be explained by the relatively short overlap sequence that uracil excision requires (typically 7–12 nucleotides [9], compared to, e.g., 12 nucleotides for PTO-based cloning [5] and 40 nucleotides for Gibson Assembly [10]). Theoretically, DNA fragments with cohesive ends should recombine with the same efficiency independently of how the single-stranded ends were generated. However, the protocol, purity, and quality of the DNA overhangs make all the difference. The quality and yield of synthetic oligonucleotides is typically low when approaching a size of 100 nucleobases [11]. Therefore, PCR-based assembly technologies that use short oligonucleotides are probably less error prone and more efficient. Moreover, short functional elements, such as promoters or ribosome binding sites, can easily be correctly incorporated directly in oligonucleotides that are assembled using short overlap sequences, because the total length of the oligonucleotide is kept relatively short. In our experience, 5′ "tails" (sequence added at the 5′ end of the oligonucleotides that do not anneal to the template DNA in the first PCR cycles) up to more than 100 nucleotides are possible, but often negatively affect the PCR yield.

Another way to ensure oligonucleotide quality is to build a molecular cloning pipeline that reuses benchmarked oligonucleotides. This was recently demonstrated for the uracil excision assembly and engineering of a six-gene biosynthetic pathway for porphyrin production [12] and a seven-gene heterologous pathway for production of a diterpene in *Escherichia coli* [13]. This type of standardization perfectly fits large collaborative efforts, much like BioBricks in the global iGEM project [14], and reuse of parts also enables better comparison of data.

Protocols for simple and seamless assembly of PCR products (also known as USER fusion), and the corresponding primer design, have been described and reviewed previously [15, 16]. Here, we provide protocols for simple manipulations and more complex assembly pipelines, including site-directed mutagenesis, multigene assembly, one-step cloning, and genome integration with uracil excision, and for a standardized, BioBrick uracil excision-based DNA editing pipeline.

2 Materials

2.1 Strains, Media, and Antibiotic Selection

1. Bacterial strains: *E. coli* strain NEB5α (New England Biolabs, Ipswich, USA) is used as a cloning host. *E. coli* BL21, K12 MG1655, and KRX (Promega, Madison, USA) are used for uracil excision combined with genomic integration (see below).
2. Growth media: SOC (20 g Bacto-Tryptone, 5 g yeast extract, 10 mM NaCl, 2.5 mM KCl, 20 mM $MgSO_4$, 20 mM glucose, water up to 1 L), 2×YT (16 g Bacto-Tryptone, 10 g yeast extract, 5 g NaCl, water up to 1 L), and LB (10 g Bacto-Tryptone, 5 g yeast extract, 10 g NaCl, water up to 1 L) (all reagents can be purchased from Sigma-Aldrich, St. Louis, USA).
3. Antibiotics: chloramphenicol (25 μg/mL), kanamycin (50 μg/mL), and tetracycline (50 μg/mL) (Sigma-Aldrich, St. Louis, USA). For clonetegration, half concentration is used with all antibiotics.

2.2 PCR Components

1. DNA polymerase: uracil excision-compatible PCR products are amplified using the proofreading PfuX7 DNA polymerase [17] (*see* **Note 1**). Cloned Pfu DNA Polymerase Buffer (Agilent Technologies, Santa Clara, USA) is used to buffer the reaction mixture.
2. Oligonucleotides (Integrated DNA Technologies, Inc., Coralville, USA) are designed with melting temperatures (T_m) of ca. 60°C. Additionally, all oligonucleotides contain one uracil, typically placed 7–12 nucleotides from the 5′ end (*see* **Note 2**). Upon uracil excision, the generated single-stranded ends should have melting temperatures between 10 and 30°C [18] (*see* **Note 3**).
3. Template DNA: plasmid DNA is isolated using the NucleoSpin® Plasmid QuickPure Kit (Macherey-Nagel, Bethlehem, USA). Plasmid aliquots are kept at −20°C (*see* **Note 4**).
4. PCR purification: PCR products are purified using a PureLink™ Quick Gel Extraction and PCR Purification Combo Kit (Thermo Fisher Scientific Inc., Waltham, USA).
5. Template DNA removal: *DpnI* (20,000 U/mL) (New England Biolabs, Ipswich, USA) is used to degrade methylated template DNA after the PCR.

2.3 USER Cloning

1. USER™ enzyme mix (New England Biolabs, Ipswich, USA).
2. USER reaction is performed in 5× Phusion HF Buffer (Life Technologies, Grand Island, USA) or Cloned Pfu DNA Polymerase Buffer (Agilent Technologies, Santa Clara, USA).

2.4 Plasmid DNA

1. Vectors: a series of pOSIP vectors is described in St-Pierre et al. [19] and can be obtained from Addgene (Addgene, Cambridge, USA). Duet vectors are available from Merck Millipore (EMD Millipore, Billerica, USA) or Addgene.

3 Methods

The protocols described here showcase the versatility of the uracil excision methodology and include protocols for (1) simple introductions of mutations, deletions, and insertions in DNA, (2) multigene assembly, (3) direct assembly and genome integration, and (4) using standardized BioBricks for assembly of pathways. The first uracil excision protocol describes the introduction of mutations, insertions, or deletions by one-fragment whole-plasmid synthesis and is largely based on the overall principles described by Nørholm [17]. Multigene assembly is performed as described previously [20] with some modifications. The third uracil excision protocol adds direct genome integration (clonetegration [19]) to the uracil excision portfolio. The optimal design parameters for multigene assembly and uracil excision combined with clonetegration have recently been explored [18]. Detailed information on clonetegration including vectors and an oligonucleotide list for colony PCR is described in St-Pierre et al. [19]. The fourth uracil excision protocol describes two operations of a fully standardized assembly procedure. The first standardized operation encompasses cloning of genes of interest into an entry vector using gene-specific oligonucleotides with fixed extensions mediating cloning. This vector contains all elements required for protein production in *E. coli* and can therefore be used straightaway for monitoring proper transcription and translation. The second standardized operation is assembly of entry fragments into multigene constructs using pairs of oligonucleotides with generic annealing parts, but distinct cloning mediating extensions. These oligonucleotides facilitate directional and specific assembly of any number of fragments. For detailed description of the options and limitations of such a standardized design, please refer to Nielsen et al. [12].

3.1 PCRs

The PCRs are performed using 1 μL PfuX7 DNA polymerase (the optimal concentration is typically batch dependent and should be empirically determined when purifying the polymerase – after desalting of his-tagged-purified PfuX7 [17], we typically determine the optimal concentration by titrating the amount of PfuX7 in a standard PCR reaction), 5 μL 10× Cloned Pfu Polymerase Buffer, 5 μL dNTP mix (25 mM each of dATP, dTTP, dGTP, dCTP), 2 μL DNA template (150 ng μL), 5 μL forward primer (5 μM), 5 μL reverse primer (5 μM), 1.2 μL $MgCl_2$ (50 mM) (it may be advantageous to optimize the $MgCl_2$ concentration from batch to batch PfuX7 by titrating the final concentration from 1 to 5 mM), and

29.8 μL nuclease-free water. The PCR involves an initial denaturation step at 98°C for 2 min, then 20 cycles of 98°C for 20 s, 58°C for 20 s, and 72°C for 45 s/kbp. Finally, the thermocycler is programmed for 72°C for 8 min and stored at 12°C.

3.2 Analysis and Purification of PCR Results

PCR products are analyzed by standard agarose gel electrophoresis. The resulting PCR products may be purified using any PCR cleanup kit.

3.3 Simple Protocol for Site-Directed Mutagenesis, Insertions, or Deletions

Mutations, deletions, or insertions in plasmid constructs are made by amplifying the whole plasmid with uracil-containing oligonucleotides that incorporate these new features. The extraordinarily simple protocol involves adding USER™ enzyme mix and *DpnI* directly to the PCR reaction mix described above; incubate for 1 h at 37°C and 20 min at 16°C in a thermocycler followed by direct transformation of 3 μL of the reaction mixture into 17 μL chemically competent cells (*see* Sect. 3.5). Oligonucleotide design is very flexible, but general guidelines can be found in Sect. 2.2, and it is recommended to try software-assisted design tools such as AMUSER [21].

3.4 Uracil Excision-Assisted Multigene Assembly

3.4.1 Simple Multigene Assembly with Non-purified Fragments

For assembly of two or more fragments, equal volumes of each PCR reaction are mixed in a total volume of 10 μL and buffered using the 5× Phusion HF Buffer (*see* **Note 5**). For template removal, *DpnI* is added prior to USER™ enzyme mix and incubated for 1 h at 37°C. The *DpnI* enzyme is deactivated by incubation at 65°C for 10 min. After 5 min on ice, 1 μL of USER™ enzyme mix is added to the reaction tubes, and uracil excision is accomplished by incubating the sample at 37°C for 15 min. Subsequently, DNA assembly is executed by cooling down the reaction to below the melting temperature of the cohesive ends for at least 15 min.

3.4.2 Multigene Assembly with Purified Fragments

Purified DNA fragments (*see* **Note 6**) are assembled as described for the non-purified DNA fragments except that 100 ng of each fragment is used and the *DpnI*-assisted template elimination step can be omitted.

3.5 Chemical Transformation of *E. coli* NEB5α Cells

17 μL of chemically competent *E. coli* NEB5α cells are mixed with 3 μL of the assembly mix described above and incubated for 15 min on ice followed by a heat shock at 42°C for 1 min (*see* **Note 7**). Following the heat shock, 1 mL of LB medium is added, and the cells are incubated for 1 h at 37°C, followed by plating on solid LB medium with the appropriate antibiotic selection for 16 h at 37°C. For selection with antibiotics like ampicillin or carbenicillin, the cells can be spread without a 1 h recovery step.

3.6 One-Step Uracil Excision Assembly and Genome Integration

Amplify one of the pOSIP backbones (*see* **Note 8** and [19]) with the oligonucleotides 5′-AGATGCAUGGCGCCTAACC-3′ and 5′-AGCCCTCUAGAGGATCCCCGGGTAC-3′ and the DNA to be integrated on the genome with 5′-AGAGGGCU-3′ followed by a gene-specific forward annealing sequence and 5′-ATGCATCU-3′ followed by a gene-specific reverse annealing sequence using the PCR conditions described above. Gel purify the amplified DNA, and make an assembly mix as described above except for using a molar ratio of 3:1 between insert and vector. Transform *E. coli* cells as described above. Recover the cells in SOC medium at 37°C for 1 h, spread the cells on LB agar plate containing the appropriate antibiotic, and incubate the plate at 30°C for 20 h. Perform a standard colony PCR to confirm the clones are integrated as described in St-Pierre et al. [19].

3.7 Standardized BioBrick Bioengineering Pipeline with Uracil Excision

Make initial entry clones by PCR amplifying the pET-Duet-1 vector using the oligonucleotides 5′-AGCACTGGUCATTGCTA ATGCTTAAGTCGAACAG-3′ and 5′-ACCACTGGUCATTGC TTATCTCCTTCTTAAAGT-3′ (*see* **Note 9**). PCR amplify gene-coding sequences with 5′-ACCAGTGGU-3′ followed by a gene-specific forward annealing sequence and 5′-ACCAGTGCU-3′ followed by a gene-specific reverse annealing sequence. In the standardized entry clones, 5′-ATGACCAGTGGT-3′ that translates into MTSG is added to the 5′ end, and 5′-AGCACTGGTCA TTGC-3′ that translates into TSGHC is added to the open reading frame. Make sure that the oligonucleotides anneal in frame with the coding sequence. At this stage, genes of interest can be tested for proper transcription and translation using selective ^{35}S-methionine labeling of gene products in the presence of rifampicin (*see* **Note 10** and [22]). The standardized 5′ end may facilitate a more predictable translational initiation rate, as previously described for similar translational fusions [23, 24], and the standardized 5′ and 3′ sequences serve as anneal sites for collections of standardized oligonucleotides for higher-order assemblies, independent of the specific genes inserted in the entry vectors. Higher-order assemblies are generated with oligonucleotides with the same overall design: linker + control element + annealing sequence. When generating the pET-Duet-1-based entry vector as described above, the forward annealing sequence for downstream multigene assembly is 5′-ATAAGCAAT GACCAGTGGT-3′, and the reverse annealing sequence is 5′-TAATGTAAGTTAGCTCACTCATTAG-3′. The principle is schematically illustrated in Fig. 2. The setup will allow the buildup of a library of benchmarked oligonucleotides where differently designed linkers have been validated for correct assembly. Examples of validated linkers are 5′-ACACCGACU-3′/5′-AGTCGGTGU-3′, 5′-ACGCTGCTU-3′/5′-AAGCAGCGU-3′, 5′-AGACGTCAU-3′/5′-ATGACGTCU-3′, 5′-AGGTCTGAGU-3′/5′-ACTCAGACCU-3′, 5′-ATAGGCTTU-3′/5′-AAAGCCTAU-3′, and 5′-AACGTGGAU-3′/5′-ATCCACGTU-3′ [12, 13]. Examples of control elements are

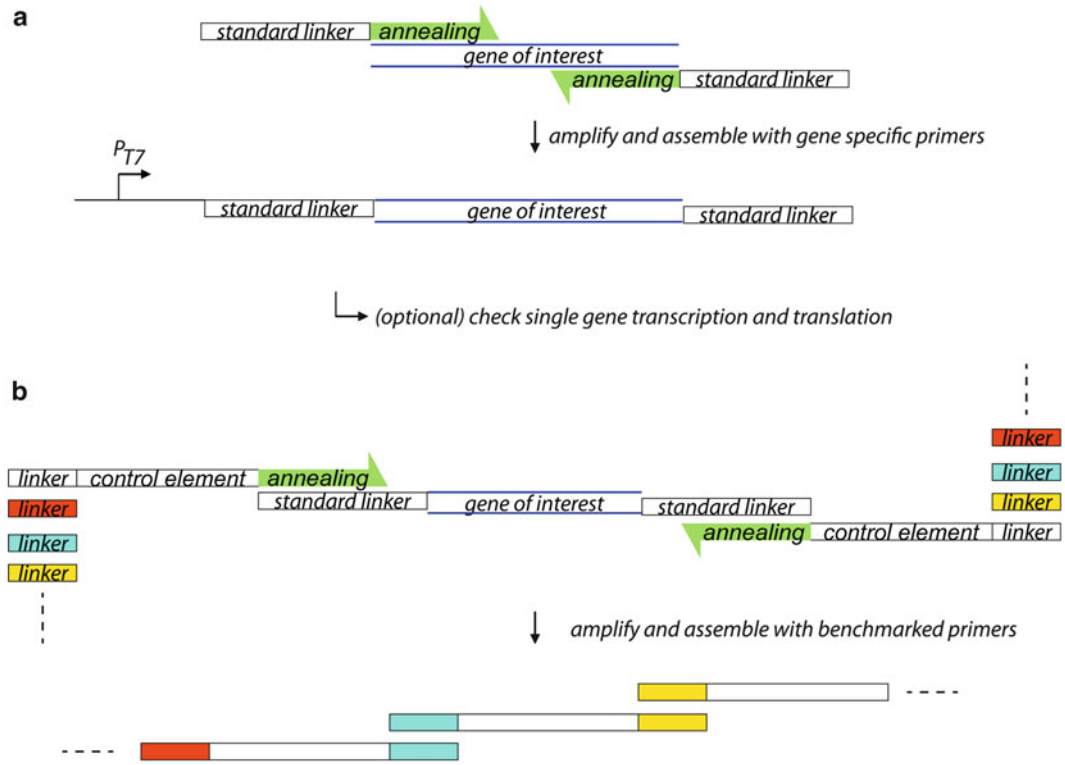


Fig. 2 Illustration of the two-step, uracil excision-based, standardized pipeline for multigene engineering. (**a**) In the first step, genes of interest are cloned with standardized linkers into an entry vector. In the entry vector, an orthogonal T7 phage promoter allows for assessment of proper transcription and translation by ^{35}S-methionine labeling in the presence of rifampicin (rifampicin blocks transcription of endogenous genes by inhibiting the endogenous *E. coli* RNA polymerase). (**b**) The standardized linkers allow the use of standardized oligonucleotides for re-amplification and construction of multigene constructs with benchmarked linkers and functional elements such as promoters and ribosome binding sites. Linkers for uracil excision are relatively short, thus allowing for larger control elements to be incorporated in standard oligonucleotides

constitutive promoters such as P_{trc} followed by randomized Shine-Dalgarno sequences (for details, *see* 12] and the phage promoter P_{T7} followed by the lac operator and a consensus Shine-Dalgarno sequence (for details, *see* 13]. The protocols for assembly are as described above (*see* **Note 11**).

4 Notes

1. Commercially available proofreading DNA polymerases with similar characteristics are available as Phusion U Hot Start DNA Polymerase (Thermo Fischer Scientific, Pittsburgh, USA) and KAPA HiFi Uracil+ (Kapa Biosystems, Inc., Wilmington, USA).

2. Oligonucleotides can be designed using the PHUSER or AMUSER software [21, 25].
3. The melting temperature of the overhangs can be calculated by online software tools such as the Tm calculator from Thermo Fischer Scientific.
4. Plasmid aliquots should contain a small volume (max. 50 μL) to avoid repeated cycles of freeze thawing.
5. According to the supplier (New England Biolabs, Ipswich, USA), the USER™ enzyme is active in all standard reaction buffers. We routinely use buffers such as Phusion HF, NEB4, cloned Pfu buffer, and T4 ligase buffer.
6. In our experience, purification in some cases enhances the efficiency and fidelity of the assembly reaction, possibly due the removal of interfering oligonucleotides [18], but it also complicates the protocol.
7. We routinely use between 30 s and 2 min for heat shock – the optimal incubation time depends on the plasticware and the heat block and can be optimized empirically.
8. Clonetegration is highly dependent on the kind of integrase in the pOSIP vector and the efficiency of the competent cells. Before you select the strain and vector for integration, check if the strain contains the *attB* site in the genome corresponding to the integrase and *attP* site in the vector. For example, in the case of pOSIP-KO (containing phage 186 integrase), MG1655 contains two corresponding *attB* sites, whereas BL21 (DE3) contains only one.
9. The protocol is described for uracil excision cloning, since this is the technique most often applied in our lab. The concept and principles of standardized assembly, however, are by no means limited to this cloning technique. On the contrary, the principles can be implemented with any PCR-based cloning technique as well as several restriction enzyme-based techniques as described in Nielsen et al. [12].
10. While his technique should be applicable to all *E. coli* strains containing T7-RNA polymerase, it is our experience that BL-21 (DE3) is superior regarding the 35-S labeling of proteins. We cannot say whether this is attributed to increased uptake and incorporation of labeled methionine, efficiency of cell lysis, or another parameter, but in side-by-side comparisons, BL-21 (DE3) consistently gives us the strongest labeling signals. Any defined media can be used, but we have found that the PASM-51 media developed by Studier (2005) yields robust expression of many different protein types in various *E. coli* expression

strains. By depleting the media of methionine, more efficient labeling is achieved.

11. The described oligonucleotides facilitate directional and specific assembly of any number of fragments, although efficiency decreases as the number of fragments increases. In our lab, 3–5 fragments (including the vector backbone) are routinely assembled using this protocol.

References

1. Mullis KB, Faloona FA (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 155:335–350
2. 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
3. Smith C, Day PJ, Walker MR (1993) Generation of cohesive ends on PCR products by UDG-mediated excision of dU, and application for cloning into restriction digest-linearized vectors. PCR Methods Appl 2:328–332
4. Nour-Eldin HH, Hansen BG, Nørholm MHH et al (2006) Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. Nucleic Acids Res 34:e122
5. Blanusa M, Schenk A, Sadeghi H et al (2010) Phosphorothioate-based ligase-independent gene cloning (PLICing): an enzyme-free and sequence-independent cloning method. Anal Biochem 406:141–146
6. Tillett D (1999) Enzyme-free cloning: a rapid method to clone PCR products independent of vector restriction enzyme sites. Nucleic Acids Res 27:26e
7. Aslanidis C, de Jong PJ (1990) Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res 18:6069–6074
8. Gibson DG, Young L, Chuang R-Y et al (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Publ Group 6:343–345
9. Bitinaite J, Nichols NM (2001) DNA cloning and engineering by uracil excision. Wiley, Hoboken
10. Gibson DG (2011) Enzymatic assembly of overlapping DNA fragments. Methods Enzymol 498:349–361
11. LeProust EM, Peck BJ, Spirin K et al (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res 38:2522–2540
12. Nielsen MT, Madsen KM, Seppälä S et al (2014) Assembly of highly standardized gene fragments for high-level production of porphyrins in *E. coli*. ACS Synt Biol 4 (3):274–282
13. Nielsen MT, Ranberg JA, Christensen U et al (2014) Microbial synthesis of the forskolin precursor manoyl oxide in enantiomerically pure form. Appl Environ Microbiol 80 (23):7258–7265
14. Shetty RP, Endy D, Knight TF Jr (2008) Engineering BioBrick vectors from BioBrick parts. J Biol Eng 2(5)
15. Salomonsen B, Mortensen UH, Halkier BA (2014) USER-derived cloning methods and their primer design. Methods Mol Biol (Clifton, NJ) 1116:59–72
16. Nour-Eldin HH, Geu-Flores F, Halkier BA (2010) USER cloning and USER fusion: the ideal cloning techniques for small and big laboratories. Methods Mol Biol (Clifton, NJ) 643:185–200
17. Nørholm MHH (2010) A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. BMC Biotechnol 10:21
18. Cavaleiro AM, Kim SH, Seppälä S et al (2015) Accurate DNA assembly and genome engineering with optimized uracil excision cloning. ACS Synth Biol. doi:10.1021/acssynbio.5b00113
19. St-Pierre F, Cui L, Priest DG et al (2013) One-step cloning and chromosomal integration of DNA. ACS Synt Biol 2(9):537–541
20. Geu-Flores F, Nour-Eldin HH, Nielsen MT et al (2007) USER fusion: a rapid and efficient method for simultaneous fusion and cloning of multiple PCR products. Nucl Acids Res 35, e55
21. Genee HJ, Bonde MT, Bagger FO et al (2014) Software-supported USER cloning strategies for site-directed mutagenesis and DNA assembly. ACS Synt Biol 4(3):342–349

22. Nevin DE, Pratt JM (1990) A coupled in vitro transcription-translation system for the exclusive synthesis of polypeptides from the T7-promoter. FEBS Lett 291:259–263
23. Kudla G, Murray AW, Tollervey D et al (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. Science (New York, NY). 324:255–258
24. Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. Science (New York, N.Y.). 342, 475–479
25. Olsen LR, Hansen NB, Bonde MT et al (2011) PHUSER (Primer Help for USER): a novel tool for USER fusion primer design. Nucleic Acids Res 39:W61–W67

Quantitative Physiology Approaches to Understand and Optimize Reducing Power Availability in Environmental Bacteria

Pablo I. Nikel and Max Chavarría

Abstract

The understanding of how carbon fluxes are distributed through a metabolic network offers an overview of the pathways that a given microorganism uses to produce energy, reducing power, and biomass. These invaluable data are related to the physiological state of the cell and provide information about the metabolic potential of microorganisms for specific environmental and biotechnological applications such as the degradation of toxic compounds (e.g., hydrocarbons) or the targeted production of high value-added products (e.g., lipids). Here, we propose a general approach to explore the pathways involved in NADPH balance in bacteria, which are in turn responsible for maintaining redox homeostasis and endowing the microorganism with the ability to counteract oxidative stress. We focus on the fluxes catalyzed by $NADP^+$-dependent enzymes in the metabolic network of the model soil bacterium *Pseudomonas putida* KT2440. This environmental microorganism is a promising cell factory for a number of NADPH-dependent biotransformations, including industrial and bioremediation processes. The relevant enzymes involved in redox balance in strain KT2440 are (1) glucose-6-phosphate dehydrogenase, (2) 6-phosphogluconate dehydrogenase, (3) isocitrate dehydrogenase, (4) malic enzyme, and (5) 2-keto-6-phosphogluconate reductase. NADPH can be generated or consumed by other enzymatic reactions depending on the microorganism; however, the first four enzymes listed above are recognized as a major source of reducing power in a wide variety of microorganisms. The present protocol includes a first stage in which the NADPH balance is derived from fluxomic data and in vitro enzymatic assays. A second step is then proposed, where the redox ratios of pyridine dinucleotides and the cell capacity to counter oxidative stress are qualitatively correlated.

Keywords: Central carbon metabolism, Fluxomics, Metabolic optimality, NADPH, Oxidative stress, *Pseudomonas putida*

1 Introduction

1.1 Fluxomics as a Tool for Metabolic Studies

The emergence of systems biology as a field has led to the development of a large number of analytical techniques to analyze the different levels of cellular organization in many organisms [1]. The high-throughput tools (the so-called "omic" techniques) currently available include those for the analysis of (1) genome

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 39–70,
DOI 10.1007/8623_2015_84, © Springer-Verlag Berlin Heidelberg 2015, Published online: 04 June 2015

(genomics), (2) gene transcription (transcriptomics), (3) protein abundance (proteomics), and (4) metabolite profiles (metabolomics) [2–4]. Each of these data sets provides snapshots of the components and processes present in the cell at a given time. With the comprehensive view of the cellular processes that systems biology brings forth, all this information is integrated to gain a more holistic view of the overall cellular performance [5, 6]. Fluxomics, which encompasses metabolic flux analysis (MFA), integrates all this information simultaneously and aims at measuring the in vivo pathway activity or enzyme reaction rates (i.e., fluxes) in the central carbon metabolism [7]. The metabolic flux distribution is the final result of the interplay of gene expression, protein and metabolite concentrations, and regulation at the level of enzymatic activity, i.e., it represents the metabolic phenotype [8]. Therefore, this technique helps representing the concept of systems biology as it reflects the total system rather than its individual parts [9].

Information on the intracellular metabolic flux distribution is of major importance in several fields of microbiology. MFA allows identifying routes used by an organism to degrade a particular carbon source [10–15], the response of cells to changes in the environmental conditions [16], or more specific information such as the pattern of regulation of enzymatic activities in central carbon metabolism in vivo [17]. Depending on the experimental design, it is possible to identify the function of a protein in catalyzing a metabolic reaction or its regulation on central metabolism (e.g., transcriptional regulators) by including the appropriate mutant strains [17]. Fluxomics is also important for fields such as synthetic biology because it allows to redesign metabolic pathways and to genetically modify microorganisms in a more rational way in order to obtain a desired product [18]. Some specific examples of MFA applications in the area of hydrocarbon degradation and lipid production include several studies in *Pseudomonas putida*, a soil bacterium capable of degrading both aliphatic and aromatic hydrocarbons [10, 14, 19], the evaluation of microalgae (e.g., the green alga *Chlorella protothecoides*) as cell factories to produce triacylglycerols for biodiesel production [20], and the systematic analysis of the pathways involved in the biosynthesis of the biopolymer poly(3-hydroxybutyrate) in *Cupriavidus necator* [21] and recombinant *Escherichia coli* strains [22].

How is MFA performed? Typically, an MFA experiment includes both an experimental stage and a bioinformatic approach. MFA is carried out by applying mass balances on steady-state metabolic models. This often results in an underdetermined system of linear equations that require other data (such as extracellular fluxes) to be solved [11, 13, 23–27]. Thus, three aspects must be considered to perform MFA experiments: (1) measurements of extracellular rates (e.g., carbon source consumption or production of some metabolites), (2) ^{13}C-labeling experiments to follow the pattern of carbon source distribution through the central

metabolism, and (3) a stoichiometric model for the metabolism of interest. Below we describe the procedures to obtain each of these parameters.

1.1.1 Measurements of Extracellular Rates

The measurement of extracellular fluxes is required to be integrated with ^{13}C-labeling patterns (explained in Sect. 1.1.2) within a stoichiometric model to obtain intracellular fluxes. Usually, extracellular rates include the consumption of the carbon source, specific rate of (by-) product secretion, and biomass production. It is generally sufficient to determine the rate of carbon source uptake to obtain intracellular fluxes; however, the extracellular flux of a (by-) product of the central metabolism, such as acetate or ethanol in *E. coli*, is often determined as well [26, 28]. Measurements of extracellular fluxes are performed during the exponential phase of growth, where a *pseudo*-steady state is assumed.

MFA has been reported in cultures of several bacteria using glucose, fructose, xylose, and malate, among many other carbon sources [10, 11, 29, 30]. Here lies one of the limitations to perform fluxomic experiments: the availability and high cost of ^{13}C-labeled carbon sources for experiments of labeling patterns (*see* Sect. 1.1.2). In addition, the detailed knowledge on how the degradation of each carbon source occurs is often lacking for non-model bacteria. This information is essential to define the stoichiometric model, as described in Sect 1.1.3. The consumption of a given carbon source can be measured by different analytical methods, depending on its chemical nature, e.g., by high-performance liquid chromatography (HPLC) coupled to mass spectrometry (MS), or using specific enzyme-based kits coupled to ultraviolet (UV) spectroscopy or chemi-/bioluminescence. As stated above, most fluxomic methods have been performed using glucose as the carbon source [10, 11, 31, 32]. ^{13}C-Labeled glucose is available in different forms: uniformly labeled in all the carbon atoms ([U-^{13}C]-glucose) or in specific carbon positions (e.g., glucose labeled in position 1, [1-^{13}C]-glucose). Furthermore, glucose metabolism is widely known in a number of microorganisms, and its concentration is easily quantified either by HPLC or a commercial kit coupled to NADPH formation (measured at 340 nm) using a mixture of hexokinase (HK) and glucose-6-phosphate (G6P) dehydrogenase (G6PDH). In the present protocol, we describe a general fluxomic method using glucose as the carbon source.

1.1.2 ^{13}C-Labeling Experiments

The most important information for an MFA experiment is obtained from ^{13}C-labeling experiments [25, 33]. MFA uses stoichiometric models of metabolism and MS methodologies to elucidate the transfer of moieties containing isotopic tracers from one metabolite into another. Relevant information about the operativity of the metabolic network is thus derived from these measurements. Labeling experiments are based on several assumptions: (1) in the exponential phase of growth, the metabolism can be

considered to be in a *pseudo*-steady state, i.e., the flux turnover of a specific metabolite is several orders of magnitude larger than the changes in the concentration of this metabolite over time (note that the *pseudo*-steady-state assumption does not imply the metabolite concentration is at a fixed steady state; it just indicates that the rate of change of these metabolites is so fast that their concentrations can be adjusted to a new steady state very rapidly), (2) ^{13}C-labeled isotope effects on biochemical reaction rates are insignificant, (3) the entire knowledge of the destination of each carbon atom in the model is available or can be inferred from the obtained ^{13}C-labeling data, and (4) the specific stoichiometry of central carbon metabolism reactions is known [7, 34, 35].

In MFA experiments, the ^{13}C-labeled carbon source is fragmented by the action of enzymatic reactions within the central metabolism in a way that a significant fraction of the carbon atoms are incorporated into proteinogenic amino acids, which leads to characteristic labeling patterns when analyzed by MS methodologies [25, 33]. The labeling pattern of these molecules (generally after fragmentation into ions) provides valuable information about the metabolic origin of the various amino acids (Fig. 1). The actual fluxes within the metabolic network can be assigned by integrating the relative isotopic abundance of metabolic intermediates with stoichiometric metabolic models and experimental physiological data. For the quantification of stable-isotope-labeled proteinogenic amino acids, two main analytical methods are commonly used, (1) nuclear magnetic resonance (NMR) and (2) gas chromatography coupled to MS (GC–MS), the latter being the most widely adopted.

NMR exploits the magnetic properties of the isotopes to distinguish between them (e.g., ^{12}C versus ^{13}C). Therefore, it is possible to track the positions where ^{13}C atoms have been incorporated in the amino acids with this technique [36, 37]. Also, NMR provides valuable labeling information such as carbon–carbon or carbon–nitrogen coupling. Despite these strengths, NMR is less used for metabolic flux analysis than GC–MS. The reason is very simple: the high sensitivity of MS and rapid data generation place this technique in advantage over NMR. Just 0.5 mg of CDW are required to complete the proteinogenic amino acid analysis via GC–MS, while at least 5 mg of CDW are needed for NMR measurements [38, 39]. In GC–MS measurements, the amino acids are first separated by GC and subsequently analyzed by MS. With the mass distributions in labeled proteinogenic amino acids obtained from mass spectra, MFA is performed by deriving information on specific metabolic steps. These data are often expressed as flux ratios because they correspond to a proportion (expressed as a relative ratio of the contributions from each relevant pathway to the biosynthesis of a common intermediate). For example, in the methodology developed by Sauer and

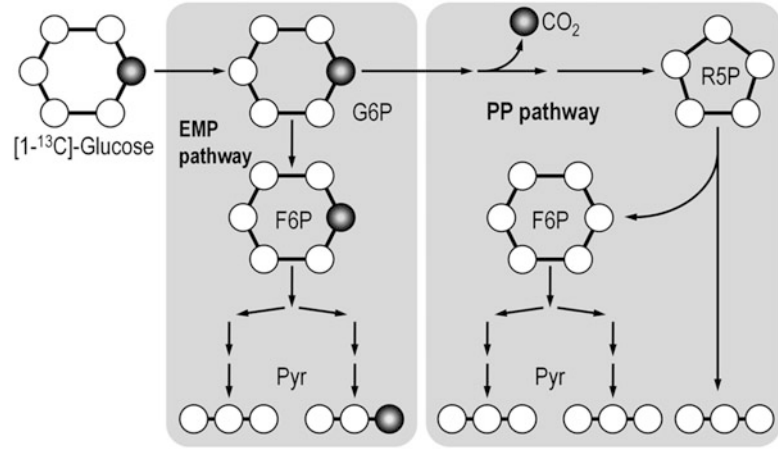


Fig. 1 Example of a typical ^{13}C-tracing experiment. Metabolic flux analysis allows for the determination of individual fluxes within a biochemical network by previously detecting the relative and positional abundance of ^{13}C in selected proteinogenic amino acids. These amino acids, in turn, come from central metabolites, which provide the link between the actual ^{13}C enrichment determination and metabolic fluxes in the metabolic network. In this example, the relative enrichment of ^{13}C in the pool of pyruvate molecules permits to identify its metabolic origin. If cells are grown on [1-^{13}C]-glucose, labeled pyruvate molecules can only stem from the linear Embden–Meyerhof–Parnas (EMP) pathway. As the oxidative decarboxylation of 6-phosphogluconate through the pentose phosphate (PP) pathway eliminates the carbon atom in the 1-C position, all pyruvate molecules originated from this metabolic sequence are unlabeled. Depending on the protocol adopted for determinations, the pattern of pyruvate labeling can be deduced from that in alanine, valine, leucine, and/or isoleucine. Note that some bioreactions have been lumped in the diagram for the sake of clarity. *G6P* glucose-6-phosphate, *F6P* fructose-6-phosphate, *R5P* ribulose-5-phosphate, *Pyr* pyruvate

collaborators [11, 13, 23, 25], a total of 14 flux ratios are obtained that include information of specific points of central metabolism. Hence, pyruvate from the Entner–Doudoroff (ED) pathway quantifies the amount of pyruvate produced through the activity of this particular pathway, while serine from Embden–Meyerhof–Parnas (EMP) pathway gives information about the activity of this linear glycolysis pathway. Subsequently, the labeling patterns in amino acids and metabolic intermediates can be integrated with the extracellular rates (as explained in Sect. 1.1.1) within a metabolic model (as detailed in Sect 1.1.3) to estimate the net fluxes through the whole central metabolism.

1.1.3 The Stoichiometric Model for Central Metabolism

A complete stoichiometric model of the central metabolism is required to obtain net fluxes (*see* Fig. 2 for an example in *P. putida* KT2440). Such model should be specific for each microorganism

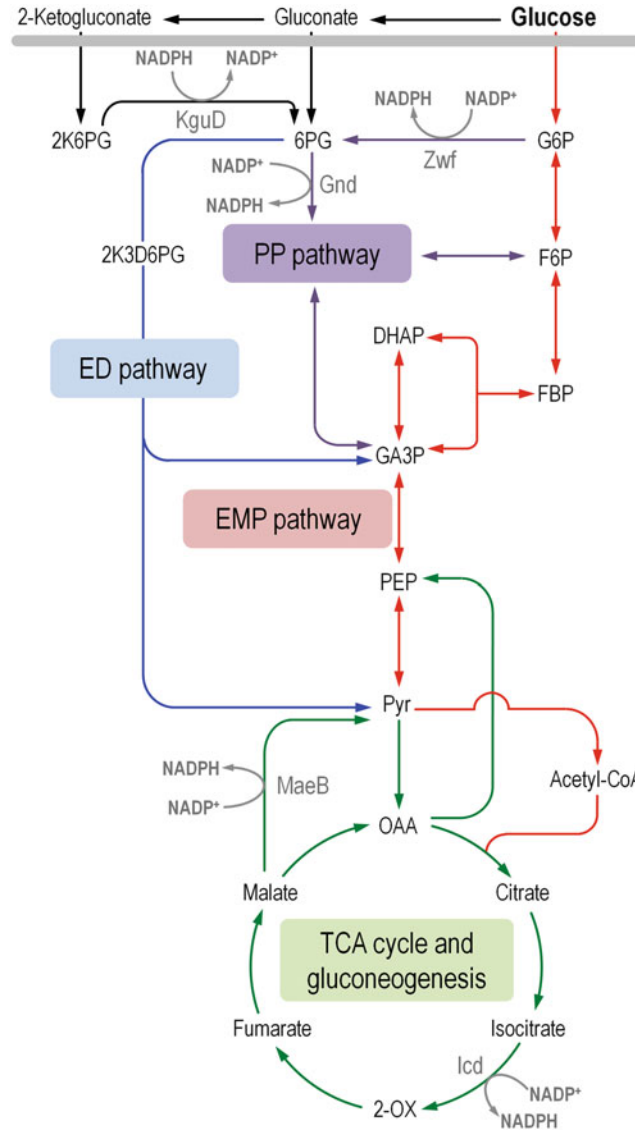


Fig. 2 Central carbon metabolism in *P. putida* KT2440. Representation of the main metabolic pathways involved in glucose catabolism in *P. putida* KT2440. Each metabolic block is highlighted in different colors. The biochemical network encompasses a set of peripheral oxidative reactions in the periplasm (*black*), the incomplete Embden–Meyerhof–Parnas (EMP) pathway (*red*), the Entner–Doudoroff (ED) pathway (*blue*), the pentose phosphate (PP) pathway (*purple*), and the tricarboxylic acid (TCA) cycle and gluconeogenesis (*green*). The enzymes responsible for the formation and consumption of NADPH are indicated in *gray*, and the *gray line* indicates the separation between periplasm and cytoplasm. Note that the oxidation of glucose to gluconate and 2-ketogluconate proceeds through the action of membrane-bound, pyrroloquinoline quinone/flavin adenine dinucleotide-dependent dehydrogenases. Some cofactors have been omitted and some bioreactions have been lumped in the diagram for the sake of clarity.

and designed by considering genomic studies (gene annotation) as well as biochemical information (if available). Unfortunately, extensive knowledge of the central metabolism including gene annotations, gene/protein functions, reaction reversibility, etc., exists only for a few model microorganisms (such as *E. coli* or yeast). A metabolic master reaction network is a requisite for MFA [11]. This is another a priori assumption of fluxomics (namely, that the components in a master biochemical network are more or less the same among different bacteria). In fact, such assumption can be considered as a limitation because the use of a general metabolic model can lead to interpretation errors [40]. However, the use of a master reaction network is widely accepted for its convenience, and the results that can be obtained from these studies are valid especially when two variables (e.g., wild-type strain versus mutant strains) or two different environmental conditions (e.g., absence/presence of stressors) are compared. For more details on metabolic models, see the protocol by Nogales entitled "Genome-scale constraint-based models" (Volume 10 in this series).

1.2 The Concept of Metabolic Optimality and Reducing Power Availability

What does *metabolic optimality* mean? An accurate definition rises from the field of computational biology: *to optimize* means to find the best solution, the best compromise among several conflicting demands subject to predefined requirements [41]. In the context of cellular performance, the "best" solution may mean maximum growth rate or highest biomass production. The justification to obtain a best solution arises from the assumption that the cell behavior adapts to the variety of conditions so that an optimal performance is ensured [41]. One of the most tightly regulated phenotypic traits in a cell is the redox balance [42]. The availability of NAD(P)H and $NAD(P)^+$ constitutes an important parameter that defines the redox homeostasis of bacterial cells, a physiological trait which can in turn be manipulated for practical purposes [43]. In particular, the catabolic formation of NADPH must be balanced with the demand of the >300 bioreactions that constitute anabolism [42].

In an attempt to explore (and manipulate) redox homeostasis in environmental bacteria, in the present protocol we focus on the quantitative analysis of NADPH formation and consumption, i.e., the regeneration of anabolic reducing power in the cell. NADH and NADPH are the two reduced nicotinamide nucleotides that constitute the basis of life [44]. While NADH provides ATP in all

Fig. 2 (continued) *FBP* fructose-1,6-bisphosphate, *G6P* glucose-6-phosphate, *F6P* fructose-6-phosphate, *6PG* 6-phosphogluconate, *2K6PG* 2-keto-6-phosphogluconate, *2K3D6PG* 2-keto-3-deoxy-6-phosphogluconate, *Pyr* pyruvate, *GA3P* glyceraldehyde-3-phosphate, *DHAP* dihydroxyacetone phosphate, *PEP* phosphoenolpyruvate, *CoA* coenzyme A, *OAA* oxaloacetate, *2-OX* 2-oxoglutarate, *Zwf* glucose-6-phosphate dehydrogenase (represented by Zwf-1, Zwf-2, and Zwf-3), *Gnd* 6-phosphogluconate dehydrogenase, *MaeB* malic enzyme, *Icd* isocitrate dehydrogenase, *KguD* 2-keto-6-phosphogluconate reductase

aerobic organisms via the process of oxidative phosphorylation, NADPH helps nullify oxidative stress and constitutes the electron donor for anabolic bioreactions [44]. For research fields such as metabolic engineering or synthetic biology, it is relevant to study how bacteria maintain an intracellular reducing environment and how fluxes are rearranged for the formation of NADPH as a source of reducing power. Several metabolic engineering strategies have focused on manipulating the NADH/NADPH metabolism in bacteria such as *E. coli* [45–47].

Yet, how does NADPH help maintaining the redox balance within the cell? Catalase, superoxide dismutase, and glutathione peroxidase are enzymes that counteract oxidative stress during aerobic respiration [48, 49]. The effectiveness of these detoxifying enzymes to fight again reactive oxygen species (ROS) largely depends on the availability of NADPH. This nucleotide supplies the reducing power necessary to suppress the oxidative damage caused by ROS [49, 50]. The resistance mechanism through the activity of glutathione peroxidase is particularly important in bacteria. Glutathione peroxidase catalyzes the reaction $2GSH + H_2O_2 \rightarrow GS–SG + 2H_2O$, where GSH represents reduced monomeric glutathione, and GS–SG represents oxidized glutathione (i.e., two GSH molecules linked by a disulfide bridge). In this process, the key metabolite is GSH, which is found at very high concentrations in several microorganisms [44, 49, 51]. This thiol maintains a strong reducing environment in the cell, and its reduced form is maintained by glutathione reductase using NADPH as a source of reducing power ($GS–SG + NADPH + H^+ \rightarrow 2GSH + NADP^+$). Thus, the reducing potential of the cell is highly dependent of NADPH availability and the production of this reducing agent is an integral part of the microbial metabolic machinery.

1.3 Central Carbon Metabolism and NADPH Regeneration

What is central carbon metabolism? Central carbon metabolism comprises all the pathways needed for transport and oxidation of a given substrate for the generation of energy and the formation of the metabolic precursors for biosynthesis of the building blocks that are in turn polymerized to form the essential cellular constituents [52, 53]. In most bacteria, central carbon metabolism basically encompasses the EMP pathway, the pentose phosphate (PP) pathway, the ED pathway, and the tricarboxylic acid (TCA) cycle (Fig. 2) [52, 54]. Interestingly, each of these pathways can be replaced by alternative biochemical reactions. In fact, all possible combinations of these classic, alternative, and abbreviated metabolic pathways can be found together in microorganisms [55]. Central metabolism is a source of ATP (energy) needed to perform most cell functions [53]. These bioreactions are also responsible of producing a sufficient number of moles of nucleotide cofactors and maintaining appropriate levels of GSH in its reduced form to fight oxidative stress as described above. What are the most important

Table 1
Biochemical reactions involved in NADPH formation and consumption in *Pseudomonas putida* KT2440

Reaction	Enzyme(s)
Glucose-6-phosphate + $NADP^+$ → 6-phosphoglucono-1,5-lactone + NADPH + H^+	Zwf-1, Zwf-2, Zwf-3, glucose-6-phosphate dehydrogenase
6-Phosphogluconate + $NADP^+$ → ribulose-5-phosphate + NADPH + CO_2	Gnd, 6-phosphogluconate dehydrogenase
D-*threo*-Isocitrate + $NADP^+$ → 2-oxoglutarate + CO_2 + NADPH + H^+	Icd, isocitrate dehydrogenase
(*S*)-Malate + $NADP^+$ → pyruvate + CO_2 + NADPH	Mae, malic enzyme
2-Keto-6-phosphogluconate + NADPH + H^+ → 6-phosphogluconate + $NADP^+$	KguD, 2-keto-6-phosphogluconate reductase

reactions for the replenishment of NADPH? It is widely recognized that, in a very generalized form, the bulk of NADPH is produced by just a few enzymatic reactions catalyzed by (1) G6PDH [42, 56, 57], (2) 6-phosphogluconate (6PG) dehydrogenase (6PGDH) [58–60], (3) $NADP^+$-dependent isocitrate dehydrogenase (Icd) [61, 62], and (4) malic enzyme (Mae) [62–64] (Table 1 and Fig. 2). Depending on the microorganism under study, NADPH formation can be mediated by other enzymatic reactions and different biochemical mechanisms. Biochemical strategies of this sort that can mediate NADPH balancing under different environmental circumstances can be divided into (1) mechanisms that avoid nucleotide imbalances in the first place and (2) biochemical processes that decouple NADPH formation from central catabolism. Imbalance-avoiding mechanisms include the appropriate choice of catabolic pathways, as observed in yeast [65], and the differential expression of isoenzymes with different cofactor specificities (e.g., the $NADP^+$-dependent acetaldehyde dehydrogenases of *Saccharomyces cerevisiae* [66] or the NAD^+- and $NADP^+$-dependent glyceraldehyde-3-phosphate dehydrogenases of *P. putida* [67]). On the other hand, the functional decoupling of catabolic NADPH formation from anabolic reactions can potentially be achieved through three distinct mechanisms: (1) the action of transhydrogenase enzymes [42, 68, 69], (2) NAD(H) kinases that can directly convert NAD(H) into NAD(P)H at the expense of ATP [70], and also (3) biochemical redox cycles, i.e., a combination of biochemical reactions or isoenzymes with different cofactor specificities which catalyzes effective transhydrogenation without affecting net catabolic fluxes [69].

Although the mechanisms listed above ensure an appropriate redox balance in several microbial cells, we decided to focus on four reactions, two of which belong to the PP pathway and two to the TCA cycle, which have been demonstrated to represent the main source of NADPH in a number of bacteria, including pseudomonads. Additionally, there are two main NADPH sinks in *P. putida* KT2440 during growth on glucose that should be considered when assessing the overall redox balance. The first (and most obvious) fate of NADPH is the anabolic buildup of biomass components. On the other hand, a considerable part of glucose is converted by most pseudomonads in organic acids, such as gluconate and 2-ketogluconate. These intermediates finally converge at the level of 6PG. One of the reducing pathways that feed this node is catalyzed by 2-keto-6-phosphogluconate reductase (KguD, Table 1 and Fig. 2). KguD uses NADPH as the cofactor to reduce 2-keto-6-phosphogluconate to 6PG (Nikel et al., unpublished), and it constitutes the second sink of NADPH in the biochemical network operated by strain KT2440 when cells grow on glucose.

G6PDH and 6PGDH are widely distributed enzymes, from bacteria to humans, and they constitute the nonreversible entry point to the PP pathway (*see* Fig. 2) [14, 71]. Particularly, during growth on hexoses, the PP pathway represents a major source of pentose phosphates for nucleotide biosynthesis and NADPH through these two consecutive $NADP^+$-dependent dehydrogenases. Carbon fluxes through the PP pathway can be different between species [11] and, depending on the demands of reducing power, these reactions may be critical to the regeneration of NADPH. On the other hand, Icd is an enzyme of the TCA cycle that catalyzes the oxidative decarboxylation of isocitrate, producing 2-oxoglutarate and CO_2 (Table 1) [61]. This process involves oxidation of isocitrate to oxalosuccinate, followed by the decarboxylation of the β-carboxyl group to the ketone, forming 2-oxoglutarate. Icd has been reported in all domains of life, and both NAD^+- and $NADP^+$-dependent enzymes have been described [72]. Finally, Mae converts malate, an intermediate of the TCA cycle, into pyruvate (Table 1 and Fig. 2), which is the end product of glycolysis and a key metabolite in the split of respiratory and fermentative metabolism (e.g., in *E. coli*). The activity of Mae can be regarded as part of a metabolic shunt where NADPH is obtained at the expense of one ATP molecule consumed by pyruvate carboxylase and one NADH molecule consumed by malate dehydrogenase [63, 73].

1.4 Overview of the Procedure

Synthetic biology and metabolic engineering approaches call for the choice of suitable hosts, which can operate in a variety of environmental conditions. One of the important aspects of an optimization procedure to obtain a robust microbial cell factory is the availability of reducing power, so that the microorganisms can

satisfy their metabolic needs and counteract oxidative stress. In this protocol, we describe a method to explore the pathways that a microorganism uses to ensure proper NADPH supply and redox balance. Specifically, we seek to associate metabolic fluxes through a set of specific enzymes with the maximum capabilities of NADPH formation in the cell, i.e., metabolic optimality as a function of redox homeostasis. For this purpose, we propose a two steps protocol, which we have previously applied in physiological and metabolic studies of the soil bacterium *P. putida* KT2440 [74, 75] (Fig. 3). As mentioned before, this Gram-negative microorganism represents a model for the biodegradation of hydrocarbons. Firstly,

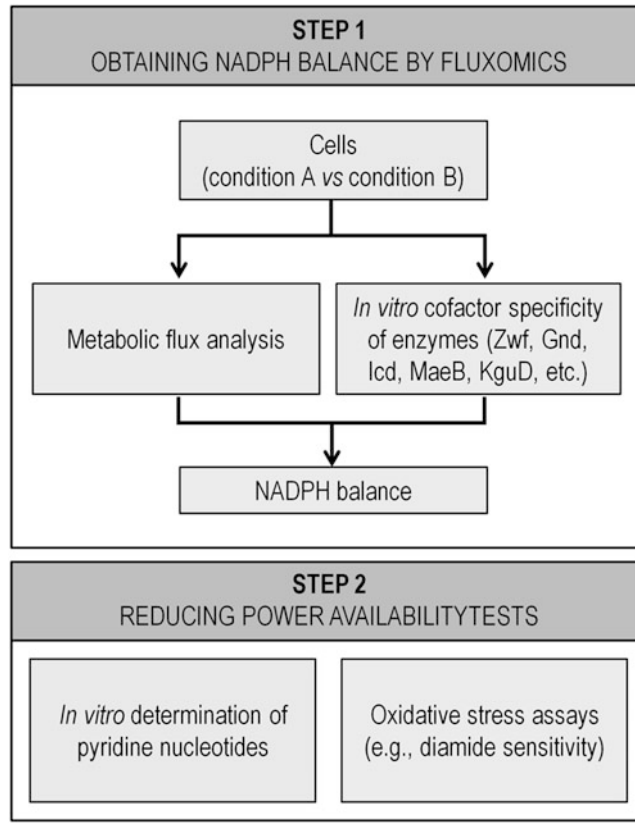


Fig. 3 Diagram of the steps proposed to study the NADPH balance in environmental bacteria. We adopted a two-step protocol, which we have applied to explore the redox homeostasis in the soil bacterium *P. putida* KT2440. Firstly, the metabolic fluxes for each of the relevant strains or environmental conditions (i.e., engineered strains, presence of stressors, alternative carbon source, etc.) are determined. The overall NADPH balance is obtained from the metabolic fluxes and the cofactor specificity of each enzyme in the biochemical network. The second step corresponds to a validation step to qualitatively associate the redox capability of the cell (i.e., redox ratios, derived from the determination of pyridine nucleotides) with its tolerance to oxidative stress

the metabolic fluxes of each of the strains or conditions in which the reducing power capability it to be determined, i.e., engineered strains, presence of stressors, culture conditions, etc., are measured. Because most fluxomic approaches have been developed to glucose-consuming cultures, in this protocol this hexose is used as carbon source. Glucose utilization has two important advantages: the metabolism of this hexose is well known in *P. putida* KT2440 (and many other microorganisms), and the labeled sugar is commercially available at a reasonable price. There are several methodologies and software platforms available to conduct MFA experiments. In Volume 10 within this series, Blank and collaborators describe a specific protocol to perform MFA experiments. In principle, any MFA protocol can be used in the approach proposed here; however, in this chapter, we will use the platform developed by Sauer and collaborators, which is well described in the literature [11, 13, 25, 26, 76, 77]. After the MFA of the cells at stake has been performed, the in vitro enzymatic assays for G6PDH, 6PGDH, $NADP^+$-dependent Icd, Mae, and KguD should be conducted. These enzymatic assays should be performed using NAD^+ and $NADP^+$ as cofactors to evaluate the specificity of each enzyme. Once the MFA data and the specificity of each enzyme are gathered, an NADPH balance is obtained from the flux distribution and the cofactor specificity of the different dehydrogenases in the biochemical network. The second step corresponds to a validation stage of the results obtained in the first step. In this step, the quantification of each cofactor is required to obtain the corresponding redox ratios and therefore to assess any possible increase or decrease in the NADPH pool according to the predictions obtained by MFA. Moreover, at this validation stage a phenotypic test is proposed to qualitatively evaluate the ability of *P. putida* KT2440 to combat oxidative stress.

2 Materials

Unless otherwise stated, all the chemicals described below were purchased from Sigma-Aldrich Co. (St. Louis, MO, USA; https://www.sigmaaldrich.com).

2.1 MFA

2.1.1 Growth Conditions and Determination of Kinetic Parameters

1. Glucose (cat. # G8270) and reagents needed for the preparation of M9 minimal medium: $Na_2HPO_4 \cdot 7H_2O$ (cat. # S9390), KH_2PO_4 (cat. # P0662), NH_4Cl (cat. # 254134), and NaCl (cat. # S9888).
2. Glucose (HK) assay kit (cat. # GAHK20). *See* **Note 1**.
3. Deionized (DI) water; resistivity $\geq$18 MΩ cm^{-1} at 25°C.
4. Benchtop microcentrifuge (capable of reaching at least 16,000×*g*).

5. Cold centrifuge for 15-mL tubes (capable of reaching at least 4,000×*g*).
6. Centrifuge tubes (15 mL) and microcentrifuge tubes (1.5 mL).
7. Nitrocellulose filters (0.45 μm) [cat. # N0271].
8. Spectrophotometer. *See* **Note 2**.
9. Sterile serological pipettes (Corning Life Sciences Inc.; Pittsburgh, PA, USA).
10. Micropipettes and the appropriate tips.
11. 0.9% (w/v) NaCl solution prepared from solid reagent [cat. # S9888]. This solution can be stored at room temperature.
12. Gyratory shaker.
13. Lab oven.
14. Analytical balance capable to measure masses within 0.0001 g.

2.1.2 ^{13}C-Labeling Experiments

1. Isotopically labeled glucose. *See* **Note 3**.
2. Cell culture (prepared in M9 minimal medium) at metabolic steady state (*see* **Note 4**). Materials are described in Sect. 2.1.1.
3. Heating block.
4. 50- and 15-mL centrifuge tubes.
5. Microcentrifuge tubes (1.5 and 2 mL).
6. Micropipettes and the appropriate tips.
7. Vortex and benchtop microcentrifuge (capable of reaching at least 16,000×*g*).
8. Cold centrifuge for 15- and 50-mL tubes (capable of reaching at least 4,000×*g*).
9. DI water.
10. 6 M HCl solution prepared from concentrated HCl [ACS reagent, 37% (w/w), cat. # H1758]. This solution can be stored at room temperature.
11. For derivatization: anhydrous, 99.8% (w/w) *N*,*N*-dimethylformamide (cat. # D4551) and *N*-*tert*-butyldimethylsilyl-*N*-methyltrifluoroacetamide with 1% (w/v) *tert*-butyldimethylchlorosilane (cat. # 375934). This reagent should be handled and stored under anhydrous conditions.
12. For the evaporation of samples: air stream.
13. Columns (*see* **Note 5**), gases, vials, filters (for GC–MS analysis).
14. Gas chromatograph coupled to a mass spectrometer. *See* **Note 6**.

2.1.3 MS Data Analysis and Flux Calculations

1. Computer equipped with an MS analysis software (*see* **Note 7**) for integrating mass spectra and data processing.

2.2 Determination of Cofactor Specificity of NAD^+- and $NADP^+$-Dependent Enzymes

2.2.1 Preparation of Cell-Free Extracts for Enzymatic Assays

1. Cell culture (prepared in M9 minimal medium) at metabolic steady state. Materials are described in Sect. 2.1.1.
2. Cold centrifuge for 15- and 50-mL tubes (capable of reaching at least 4,000×*g*).
3. DI water.
4. 10 mM sodium phosphate buffer (pH 7.5) containing 100 mM 2-mercaptoethanol (cat. # M6250). Sodium phosphate buffer (pH 7.5) is prepared by mixing 1 M solutions of NaH_2PO_4 and Na_2HPO_4 (e.g., by mixing 8.4 mL of Na_2HPO_4 and 1.6 mL NaH_2PO_4). The mixture should be diluted to 1 L and the final pH should be 7.5 (if needed, adjust the pH with the concentrated Na_2HPO_4 or NaH_2PO_4 solutions as appropriate). Then, 2-mercaptoethanol is added up to 100 mM. This buffer can be stored for up to 1 month at 4°C.
5. Cell disruptor (e.g., Omni Ruptor 4000 ultrasonic homogenizer/cell disrupter; Omni International Inc., Kennesaw, GA, USA).
6. Commercial kit for protein determination based on the Bradford method [78] (cat. # 500–0201; Bio-Rad Labs., Hercules, CA, USA; http://www.bio-rad.com).
7. 50- and 15-mL centrifuge tubes.
8. Micropipettes and the appropriate tips.

2.2.2 Assay for G6PDH

1. 250 mM glycylglycine (Gly–Gly) pH 7.5 [Gly–Gly, >99% (w/w), cat. # G1002]. This buffer can be stored for up to 1 month at 4°C.
2. DI water.
3. 60 mM G6P. The solution was prepared from G6P monosodium salt (cat. # G7879). This solution should be stored at −20°C.
4. 20 mM β-nicotinamide adenine dinucleotide phosphate ($NADP^+$). The solution was prepared from solid reagent [≥98% (w/w), cat. # N0505]. Working solutions have to be prepared freshly in DI water.
5. 20 mM β-nicotinamide adenine dinucleotide (NAD^+). The solution was prepared from solid reagent [≥98% (w/w), cat. # N6522]. Working solutions have to be prepared freshly in DI water.
6. 300 mM $MgCl_2$. The solution was prepared from solid reagent [≥98% (w/w), anhydrous, cat. # 63063] and can be stored at room temperature.

7. G6PDH solution prepared from lyophilized enzyme from *S. cerevisiae* (200–400 mg per protein, cat. # G6378). Immediately before use, prepare a solution containing 0.3–0.6 units mL^{-1} in cold DI water.
8. 15-mL centrifuge tubes.
9. Micropipettes and the appropriate tips.
10. UV-transparent cuvettes. *See* **Note 8**.
11. Water bath (e.g., Precision general-purpose water bath; Thermo Fisher Scientific Inc., Waltham, MA, USA).

2.2.3 Assay for 6PGDH

1. 100 mM Gly–Gly, pH 7.5 (cat. # G1002). This buffer should be stored at 4°C.
2. DI water.
3. 100 mM 6PG prepared from 6PG trisodium salt [≥97% (w/w), cat. # P7877]. This solution should be stored at −20°C.
4. 60 mM $NADP^+$ [≥98% (w/w), cat. # N0505]. Working solutions have to be prepared freshly in DI water.
5. 60 mM NAD^+ [≥98% (w/w), cat. # N6522]. Working solutions have to be prepared freshly in DI water.
6. 6PGDH enzyme solution prepared from lyophilized enzyme from *S. cerevisiae* (3.0–6.0 mg per protein, cat. # P4553). Immediately before use, prepare a solution containing 1.5–3.0 units mL^{-1} in cold DI water.
7. 15-mL centrifuge tubes.
8. Micropipettes and the appropriate tips.
9. UV-transparent cuvettes.
10. Water bath.

2.2.4 Assay for Icd

There are several commercial kits for determining the isocitrate dehydrogenase activity by colorimetric techniques (Sigma-Aldrich Co., https://www.sigmaaldrich.com; Abcam, http://www.abcam.com; BioVision, http://www.biovision.com; BioAssay Systems, http://www.bioassaysys.com; or Genway Biotech Inc., http://www.genwaybio.com). There are also protocols for the analysis of Icd activity using fluorescence or UV spectroscopy. Here, we describe a standard protocol, which utilizes isocitrate as a specific substrate and measures the $NADP^+$ reduction rate at 340 nm.

1. 250 mM Gly–Gly buffer (pH 7.4) (cat. # G1002). This buffer should be stored at 4°C.
2. DI water.
3. 6.6 mM D,L-isocitric acid solution prepared from D,L-isocitric acid trisodium salt hydrate [≥93% (w/w), cat. # I1252]. This solution should be stored at −20°C.

4. 20 mM $NADP^+$ [≥98% (w/w), cat. # N0505]. Working solutions have to be prepared freshly in DI water.
5. 20 mM NAD^+ [≥98% (w/w), cat. # N6522]. Working solutions have to be prepared freshly in DI water.
6. 18 mM $MnCl_2$. Solution prepared from solid reagent [≥99% (w/w), cat. # 244589], and it can be stored at room temperature.
7. Icd enzyme solution (0.3–0.6 units mL^{-1}) prepared from commercial product (*Bacillus subtilis*, recombinant from *E. coli*, cat. # 94596). Prepare freshly in cold Gly–Gly buffer.
8. 15-mL centrifuge tubes.
9. Micropipettes and the appropriate tips.
10. UV-transparent cuvettes.
11. Water bath.

2.2.5 Assay for Mae

1. 100 mM triethanolamine · HCl buffer (pH 7.4). Prepared from solid triethanolamine hydrochloride (≥99.5%, cat. # T1502). This buffer can be stored for up to 1 month at 4°C.
2. DI water.
3. 100 mM L-malic acid solution prepared from solid reagent [95–100% (w/w), cat. # M1000]. Solution should be stored at −20°C.
4. 20 mM $NADP^+$ [≥98% (w/w), cat. # N0505]. Working solutions have to be prepared freshly in DI water.
5. 20 mM NAD^+ [≥98% (w/w), cat. # N6522]. Working solutions have to be prepared freshly in DI water.
6. 20 mM $MnCl_2$ [≥99% (w/w), cat. # 244589]. This solution can be stored at room temperature.
7. Mae solution (0.25–0.50 units mL^{-1}) prepared from recombinant enzyme from *E. coli* (also known as malic dehydrogenase, cat. # 18115). Prepare fresh in DI water.
8. 15-mL centrifuge tubes.
9. Micropipettes and the appropriate tips.
10. UV-transparent cuvettes.
11. Water bath.

2.2.6 Assay for KguD

1. 100 mM Tris · HCl buffer (pH 8). Prepared from solid tris base (≥99.9%, cat. # T5941). This buffer should be stored at 4°C.
2. DI water.
3. 40 mM 2-keto-3-deoxy-6-phosphogluconic acid lithium salt solution prepared from solid reagent [≥95% (w/w), cat. # 79156]. This solution should be stored at −20°C.

4. 20 mM NADPH [≥97% (w/w), cat. # N7505]. Working solutions have to be prepared freshly in DI water.
5. 20 mM $MgCl_2$ (cat. # 63063). This solution can be stored at room temperature.
6. 15-mL centrifuge tubes.
7. Micropipettes and the appropriate tips.
8. UV-transparent cuvettes.
9. Water bath.

2.3 Determination of Pyridine Dinucleotides

1. Cold centrifuge for 15- and 50-mL tubes (capable of reaching at least 8,500×*g*).
2. DI water.
3. Cell culture (prepared in M9 minimal medium) at metabolic steady state. Materials are described in Sect. 2.1.1.
4. Benchtop microcentrifuge (capable of at least 16,000×*g*).
5. Lab oven.
6. 50 mM sodium phosphate buffer (pH 7.5). 50 mM sodium phosphate buffer (pH 7.5) is prepared by mixing 1 M solutions of NaH_2PO_4 and Na_2HPO_4 (e.g., by mixing 42 mL of Na_2HPO_4 and 8 mL NaH_2PO_4). The mixture should be diluted to 1 L and the final pH should be 7.5 (if needed, adjust the pH with the concentrated Na_2HPO_4 or NaH_2PO_4 solutions as appropriate). This buffer can be stored for up to 3 months at room temperature.
7. 0.25 M NaOH prepared from solid reagent (cat. # 221465), and it can be stored at room temperature.
8. 0.25 M HCl [ACS reagent, 37% (w/w), cat. # H1758]. This solution can be stored at room temperature.
9. 0.1 M NaCl from solid reagent (cat. # S9888). This solution can be stored at room temperature.
10. 1 M and 120 mM bicine · NaOH (pH 8.0) prepared from solid bicine (≥99%, cat. # B3876). Dissolve the bicine and adjust the pH with 5 N NaOH. Buffer should be stored at 4°C.
11. 2.5 mM 3-(4,5-dimethylthiazole-2-yl)-2,5-diphenyltetrazolium bromide (MTT) prepared from solid MTT [98%, cat. # M2128]. Prepare freshly in DI water just before use and keep the solution protected from light.
12. 80 mM ethylenediaminetetraacetic acid (EDTA) prepared from solid reagent (ACS reagent, cat. # E9884) and can be stored at room temperature.
13. 15 mM phenazine ethosulfate from solid reagent (cat. # P4544). Prepare freshly in DI water just before use and keep the solution protected from light.

14. 12.5 mM G6P solution prepared from G6P monosodium salt (cat. # G7879). This solution should be stored at −20°C.
15. 350 units mL^{-1} of alcohol dehydrogenase (ADH) prepared from the lyophilized enzyme from *S. cerevisiae* (≥300 mg per protein, cat. # A3263). Prepare immediately before use in 120 mM bicine · NaOH (pH 8.0) and keep at 4°C.
16. 0.5 units mL^{-1} of G6PDH prepared from the lyophilized enzyme from *S. cerevisiae* (200–400 mg per protein, cat. # G4134). Prepare immediately before use in 120 mM bicine · NaOH (pH 8.0) and keep at 4°C.
17. 1.25 M ethanol prepared in DI water from absolute ethanol (ACS reagent, cat. # 459844). This solution can be stored at room temperature.
18. $NADP^+$ solution for the calibration curve prepared from solid reagent (≥98%, cat. # N0505). Prepare standards freshly in DI water.
19. NAD^+ solution for the calibration curve prepared from solid reagent (≥98%, cat. #N6522). Prepare standards freshly in DI water.
20. Nunclon MicroWell plates for automation (96 wells, with lid), flat bottom, clear (Thermo Scientific Inc.; Waltham, MA, USA).
21. SpectraMax Plus 384 microplate reader (Molecular Devices LLC.; Sunnyvale, CA, USA).
22. 50- and 15-mL centrifuge tubes.
23. Micropipettes and the appropriate tips.
24. Liquid N_2.

2.4 Tolerance to Oxidative Stress Test

1. Glucose (cat. # G8270) and the reagents needed for preparation of M9 minimal medium.
2. DI water.
3. Diamide (DA) solutions for a dose–response curve. A 1 M solution is prepared in dimethyl sulfoxide (DMSO) (cat. # D8418) from the solid reagent (cat. # D3648). Prepare freshly and protect from light exposure.
4. Nunclon MicroWell plates for automation (96 wells, with lid), flat bottom, clear.
5. SpectraMax Plus 384 microplate reader.
6. 50- and 15-mL centrifuge tubes.
7. Micropipettes and the appropriate tips.

3 Methods

3.1 MFA

3.1.1 Growth Conditions and Determination of Kinetic Parameters

1. Prepare M9 minimal medium (*see* **Note 9**) containing 20 mM glucose as the only C source.
2. Determine the maximum growth rate (μ) in aerobic batch cultures. Grow cells at 30°C with shaking at 170 rpm in 250-mL baffled Erlenmeyer flasks filled with 50 mL of M9 minimal medium supplemented with 20 mM glucose. Monitor growth spectrophotometrically (*see* **Note 2**) at a wavelength of 600 nm (OD_{600}). Results of turbidity measurements are computed during exponential growth (log-linear regression of OD_{600} versus time), and $\mu(h^{-1})$ is calculated for each condition as $\mu = [\ln(OD_{600} \text{ at } t_1) - \ln(OD_{600} \text{ at } t_0)]/(t_1 - t_0)$.
3. Determine the correlation factor (k) between CDW and OD_{600} as follows. Aerobic batch cultures are developed as detailed in the preceding section. At least seven parallel 10-mL cell suspension aliquots are harvested by fast filtration at different times during exponential growth. Cultures are filtered in pre-weighed nitrocellulose filters (0.45 μm), which are subsequently washed twice with 10 mL of 0.9% (w/v) NaCl and dried at 105°C for 24 h to constant weight (*see* **Note 10**). The value of k is determined by linear regression of OD_{600} versus CDW, k being the regression coefficient in the equation $OD_{600} = k \times CDW + b$. The parameter b is a constant.
4. Determine the biomass yield on glucose ($Y_{X/S}$; g_{CDW} g glucose^{-1}) as the coefficient of a linear regression of CDW versus consumed glucose concentration during the exponential growth phase [$OD_{600} \times k = CDW = Y_{X/S} \times$ consumed glucose concentration + c]. The parameter c is a constant. Aerobic batch cultures are grown as described above, and 1-mL cell suspension aliquots are taken during the exponential growth phase (at least seven points), the OD_{600} measured, and cells are immediately harvested by centrifugation (1 min at 15,800×g) in an Eppendorf tabletop centrifuge to sediment the biomass. Glucose concentration is determined enzymatically in the culture supernatant with a glucose kit.
5. Calculate the specific rate of glucose consumption (q_S, g glucose $g_{CDW}{}^{-1}$ h^{-1}) dividing the maximum growth rate (μ) by the biomass yield on glucose ($Y_{X/S}$).

3.1.2 ^{13}C-Labeling Experiments

1. Aerobic batch cultures are developed as detailed in the preceding section (*see* **Note 11**).
2. Harvest 5–10-mL cell suspension aliquots at mid-exponential growth phase (at an OD_{600} of about 50% of the maximal value, $OD_{600} = 0.5–0.6$) by centrifugation (15 min, 4,000×g, 4°C). Cell pellets are washed twice by resuspension in 1 mL of 0.9%

(w/v) NaCl, transferred into a 2-mL Eppendorf tube, and centrifuged in a tabletop Eppendorf centrifuge at 15,800×*g* and room temperature for 5 min. *See* **Note 12**.

3. Resuspend the washed pellet in 1.5 mL of 6 M HCl, hydrolyze for 24 h at 110°C in sealed 2-mL Eppendorf tubes, and desiccate the contents overnight in a heating block at 60°C under a constant air stream.
4. Resuspend the dried hydrolysate in 30 μL of dimethyl formamide and derivatize under gentle shaking with 30 μL of *N*-*tert*-butyldimethylsilyl-*N*-methyltrifluoroacetamide containing 1% (v/v) *tert*-butyldimethylchlorosilane at 85°C for 60 min. *Tert*-butyldimethylchlorosilane allows for an efficient derivatization of the amino acids [25] and readily silylates hydroxyl groups, thiols, primary and secondary amines, amides, and carboxyl groups. The derivatized sample is then immediately transferred into an amber crimp vial and sealed with a cap.
5. Inject 1 μL of the derivatized sample into a Series 8000 gas chromatograph combined with a MD 800 mass spectrometer (Fisons Instruments PLC, Ipswich, UK) using a split ratio of 1:20 on a SPB-1 column (*see* **Note 5**). The carrier gas flow [helium, ≥ 99.996% (v/v) purity; PanGas AG, Dagmersellen, Switzerland] is set at 2 mL min^{-1}. The initial oven temperature of 150°C is maintained for 2 min and raised to 280°C with a gradient of 10°C min^{-1}. The final temperature is maintained for 2 min, and source and interface temperatures are held at 200°C and 280°C, respectively. Ions are generated by electron impact at −70 eV (full scan ranging from $m/z = 70$ to 560, with a solvent delay of 4 min). The amino acids analyzed by GC–MS are Ala, Asp, Glu, Gly, His, Ile, Leu, Phe, Pro, Ser, Thr, Tyr, and Val for uniformly labeled substrates and Ala, Asp, Ile, Phe, Leu, Ser, Thr, Tyr, and Val for [1-^{13}C]-labeled substrates.

3.1.3 MS Data Analysis

1. Perform MS analysis through custom-developed or commercial software (*see* **Note 7**). First, identify the chromatographic peaks corresponding to the amino acids of interest. The identification is based on (1) the retention time and (2) the MS spectra of the peak. Built-in databases of the mass spectrometer can help to identify each amino acid. In case of doubt, or when working with complex samples, mixtures of pure amino acid standards or spiked samples can be run in parallel to ensure the identity of each peak. Commercially available standards of most amino acids are required for MS analysis.
2. Once a suitable chromatogram with assigned peaks is obtained, the integration of specific ions in the mass spectra of each amino acid is targeted at. This information is then used to

obtain the distribution of mass isotopomers (i.e., isotopic isomers of the same molecule, only differing in the position of ^{13}C atoms) and to quantify the relative abundance of isotope peaks through program ratio of Fiat Flux software [77]. There are other software platforms freely available that allow for data processing.

3. Obtain the following flux ratios: serine derived from the EMP pathway, pyruvate derived from the ED pathway, oxaloacetate (OAA) originating from pyruvate, phosphoenolpyruvate (PEP) originating from OAA, the lower and upper bounds of pyruvate originating from malate, and the upper bound of PEP derived from the PP pathway.
4. Calculate the net fluxes with the MATLAB-based program Netto of Fiat Flux software [77] by minimizing the sum of the weighed square residuals of the constraints from both metabolite balances and flux ratios. The metabolic model used for net-flux analysis is based on a master reaction network [11] which includes 45 reactions and 33 metabolites (*see* **Note 13**). For the calculation of net fluxes, additional information is needed: (1) the stoichiometric reaction matrix, (2) the flux ratios, (3) physiological data [i.e., maximum growth rate (μ) and specific rate of glucose consumption (q_S)], and (4) precursor requirements for biomass synthesis [77].

3.2 Determination of Cofactor Specificity of NAD^+- and $NADP^+$-Dependent Enzymes

3.2.1 Preparation of Cell-Free Extracts for Enzymatic Assays

1. Grow *P. putida* cultures as described in the preceding sections.
2. Harvest 5–25-mL cell suspension aliquots at mid-exponential growth phase (at an OD_{600} of about 50% of the maximal value, $OD_{600} = 0.5–0.6$) by centrifugation (15 min, 4,000×*g*, 4°C).
3. Resuspend the pellets in the appropriate volume of 50 mM phosphate buffer (pH 7.5) containing 100 mM 2-mercaptoethanol, to achieve an $OD_{600} = 4$.
4. Disrupt the cells by sonication in the cold (five pulses of 30 s) and spin down (30 min, 14,000×*g*, 4°C) to collect cell debris.
5. Determine the protein concentration in the supernatant (i.e., the cell-free extract) as per the Bradford protocol [78] using a commercially available kit.

3.2.2 Assay for G6PDH

1. Prepare a reaction mixture by pipetting appropriate volumes of the solutions described in Sect. 2.2.2 to obtain a concentration of 43 mM Gly–Gly, 2 mM G6P, 10 mM $MgCl_2$, and 0.7 mM $NADP^+$ or NAD^+. The final volume is adjusted with DI water. Mix and equilibrate at 30°C (*see* **Note 14**).
2. Pipette 2.90 mL of the reaction mixture into suitable cuvettes (*see* **Notes 2** and **8**). Equilibrate at 30°C. Monitor the absorbance at 340 nm (A_{340}) until a constant value is reached.

3. Start the reaction by the addition of 100 μL (*see* **Note 14**) of cell-free extract (or an appropriate dilution thereof). Immediately mix by inversion and record the increase in A_{340} for approximately 10 min (*see* **Note 15**).
4. A negative control should be performed using the same volume of reaction mixture and 100 μL of the buffer used to prepare cell extracts [i.e., 50 mM phosphate (pH 7.5) and 100 mM 2-mercaptoethanol].
5. A positive control should be performed using the same volume of reaction mixture and 100 μL of a G6PDH suspension containing 0.3–0.6 units mL^{-1} instead of cell-free extract.
6. Obtain the ΔA_{340} min^{-1} using the maximum linear rate for both the test and blank using a minimum of five points in the linear region of the curve [A_{340} versus time (min)].
7. Calculate the activity of G6PDH (in units mL^{-1}) for each cofactor (i.e., NAD^+ and $NADP^+$) as follows:

$$\text{units mL}^{-1} = \left(\left[\left(A_{340}\text{min}^{-1}\right)_{\text{sample}} - \left(A_{340}\text{min}^{-1}\right)_{\text{blank}} \right] \times 3 \times \text{DF} \right) / (6.22 \times 0.1)$$

where:

3 = total volume (in mL) of assay

DF = dilution factor (whenever used)

6.22 = millimolar extinction coefficient of NADPH at 340 nm (in mM^{-1} cm^{-1})

0.1 = volume (in mL) of cell-free extract used

3.2.3 Assay for 6PGDH

1. Prepare a reaction mixture by pipetting appropriate volumes of the solutions described in Sect. 2.2.3 to obtain a concentration of 94 mM Gly–Gly, 1.7 mM 6PG, and 2.0 mM $NADP^+$ or NAD^+. Mix and equilibrate at 30°C.
2. Pipette 2.90 mL of the reaction mixture into suitable cuvettes (*see* **Notes 2** and **8**). Equilibrate at 30°C. Monitor the A_{340} until a constant value is reached.
3. Start the reaction by the addition of 100 μL (*see* **Note 16**) of cell-free protein extract (or an appropriate dilution thereof). Immediately mix by inversion and record the increase in A_{340} for approximately 5 min.
4. A negative control should be performed using the same volume of reaction mixture and 100 μL of the buffer used to make cell extracts [i.e., 50 mM phosphate (pH 7.5) and 100 mM 2-mercaptoethanol].

5. A positive control should be performed using the same volume of reaction mixture and 100 μL of a 6PGDH suspension containing 0.03–0.06 units mL^{-1} instead of cell-free extract.
6. Obtain the ΔA_{340} min^{-1} and the 6PGDH activity (i.e., units mL^{-1}) as described in Sect. 3.2.2.

3.2.4 Assay for Icd

1. Prepare a reaction mixture by pipetting appropriate volumes of the solutions described in Sect. 2.2.4 to obtain a concentration of 67 mM Gly–Gly, 0.44 mM D,L-isocitric acid, 0.60 mM $MnCl_2$, and 1.0 mM $NADP^+$ or NAD^+. Mix and equilibrate at 30°C.
2. Pipette 2.90 mL of the reaction mixture into suitable cuvettes (*see* **Notes 2** and **8**). Equilibrate at 30°C. Monitor the A_{340} until a constant value is reached.
3. Start the reaction by the addition of 100 μL (*see* **Note 16**) of cell-free protein extract (or an appropriate dilution thereof). Immediately mix by inversion and record the increase in A_{340} for approximately 5 min.
4. A negative control should be performed using the same volume of reaction mixture and 100 μL of the buffer used to make cell extracts [i.e., 50 mM phosphate (pH 7.5) and 100 mM 2-mercaptoethanol].
5. A positive control should be performed using the same volume of reaction mixture and 100 μL of an Icd suspension 0.03–0.06 units mL^{-1} instead of cell-free extract.
6. Obtain the ΔA_{340} min^{-1} and the Icd activity (i.e., units mL^{-1}) as described in Sect. 3.2.2.

3.2.5 Assay for Mae

1. Prepare a reaction mixture by pipetting appropriate volumes of the solutions described in Sect. 2.2.4 to obtain a concentration of 67 mM triethanolamine, 3.3 mM L-malic acid, 5.0 mM $MnCl_2$, and 0.3 mM $NADP^+$ or NAD^+. Mix and equilibrate at 30°C.
2. Pipette 2.90 mL of the reaction mixture into suitable cuvettes (*see* **Notes 2** and **8**). Equilibrate at 30°C. Monitor the A_{340} until a constant value is reached.
3. Start the reaction by the addition of 100 μL (*see* **Note 16**) of cell-free protein extract (or an appropriate dilution thereof). Immediately mix by inversion and record the increase in A_{340} for approximately 5–10 min.
4. A negative control should be performed using the same volume of reaction mixture and 100 μL of the buffer used to make cell extracts [i.e., 50 mM phosphate (pH 7.5) and 100 mM 2-mercaptoethanol].

5. A positive control should be performed using the same volume of reaction mixture and 100 μL of a Mae suspension 0.025–0.050 units mL^{-1} instead of cell-free extract.
6. Obtain the ΔA_{340} min^{-1} and the Mae activity (i.e., units mL^{-1}) as described in Sect. 3.2.2.

3.2.6 Assay for KguD

1. Prepare a reaction mixture by pipetting appropriate volumes of the solutions described in Sect. 2.2.6 to obtain a concentration of 55 mM Tris · HCl (pH 8), 9.5 mM $MgCl_2$, 1.5 mM NADPH, and 4 mM 2-keto-3-deoxy-6-phosphogluconic acid. Mix and equilibrate at 30°C.
2. Pipette 2.90 mL of the reaction mixture into suitable cuvettes (*see* **Notes 2** and **8**). Equilibrate at 30°C. Monitor the A_{340} until a constant value is reached.
3. Start the reaction by the addition of 100 μL (*see* **Note 16**) of cell-free protein extract (or an appropriate dilution thereof). Immediately mix by inversion and record the increase in A_{340} for approximately 15 min.
4. A negative control should be performed using the same volume of reaction mixture and 100 μL of the buffer used to make cell extracts [i.e., 50 mM phosphate (pH 7.5) and 100 mM 2-mercaptoethanol].
5. Obtain the ΔA_{340} min^{-1} and the KguD activity (i.e., units mL^{-1}) as described in Sect. 3.2.2.

3.2.7 Determination of the NADPH Balance

1. Calculate the relative cofactor dependence (CD) of each dehydrogenase (DH) from the data obtained according to Sect. 3.2 as CD_x^{DH} = specific DH activity with cofactor *x*/total specific DH activity, where *x* is any given cofactor. For example, the relative cofactor dependence of G6PDH for $NADP^+$ (termed CD_{NADP}^{G6PDH}) is obtained as the ratio CD_{NADP} = specific G6PDH activity in the presence of $NADP^+$/total specific G6PDH activity (i.e., using $NADP^+$ and NAD^+). Note that for any given DH enzyme, $CD_{NADP} + CD_{NAD} = 1$. *See* **Note 17**.
2. Obtain the net NADPH formation rate (termed f_{NADPH}^{F}, the superscript F standing for *formation*) as $f_{NADPH}^{F} = CD_{NADP}^{G6PDH} \times v_{G6PDH} + CD_{NADP}^{6PGDH} \times v_{6PGDH} + CD_{NADP}^{Icd} \times v_{Icd} + CD_{NADP}^{Mae} \times v_{Mae}$, where *v* is the net flux through the reaction catalyzed by the corresponding enzyme. Note that both *f* and *v* have the same units, as they represent fluxes,
3. Obtain the net NADPH consumption rate (termed f_{NADPH}^{C}, the superscript C standing for *consumption*) as $f_{NADPH}^{C} = CD_{NADPH}^{KguD} \times v_{KguD}$ + NADPH requirement for biomass formation. The NADPH requirement for biomass formation is directly obtained from the corresponding flux

distribution in the MFA experiment (i.e., from the flux representing biomass generation from metabolic precursors).

4. Calculate the overall NADPH balance as $r_{NADPH} = f^{F}_{NADPH} - f^{C}_{NADPH}$. This calculation allows to identify whether the physiological state of the cells is characterized by catabolic NADPH underproduction ($r_{NADPH} < 0$) or catabolic NADPH overproduction ($r_{NADPH} > 0$).

3.3 Determination of Pyridine Dinucleotides

The protocol below is based on the cycling assay originally developed by Bernofsky and Swan [79], with the modifications described elsewhere [80–82].

1. Grow *P. putida* cultures as described in the preceding sections.
2. Harvest 1.5-mL cell suspension aliquots at mid-exponential growth phase (at an OD_{600} of about 50% of the maximal value; $OD_{600} = 0.5–0.6$) by fast centrifugation (1 min, 12,500×*g*, 4°C). Depending on the cell density, a new cell culture aliquot can be added to the sediment from the first sampling. Discard the supernatant and freeze the biomass samples by rapid immersion of the Eppendorf tubes in liquid N_2 (*see* **Note 18**). Harvest another suitably large broth sample in parallel to determine CDW concn. as described in Sect. 3.1.1.
3. Add 0.3 mL of either 0.25 M NaOH [for NAD(P)H extraction] or HCl [for $NAD(P)^+$ extraction] to the frozen biomass samples.
4. Heat the samples for 15 min at 55°C.
5. Neutralize the samples by dropwise addition of 0.3 mL of either 0.1 M HCl [for NAD(P)H extraction] or NaOH [for $NAD(P)^+$ extraction]. Add 0.1 mL of 1 M bicine · NaOH buffer (pH 8.0) to all samples to equilibrate the pH.
6. Remove cellular debris by centrifugation (5 min, 12,500×*g*, room temperature), and transfer the supernatants to clean Eppendorf tubes (*see* **Note 19**).
7. Take 5 μL of the sample and add it to a single well in 96-well microtiter plates containing 90 μL of the reaction mixture. The components in the cycling reaction mixture and their final concentrations are 120 mM bicine · NaOH (pH 8.0), 0.5 mM MTT, 4.5 mM EDTA, 4.5 mM phenazine ethosulfate, and the cognate substrate (either 200 mM ethanol or 12.5 mM G6P). After addition of the extracts to the wells containing the appropriate reaction mixture [i.e., containing ethanol for NAD(H/$^+$) determinations and G6P for NADP(H/$^+$) determinations], the plates are incubated at 30°C for 5 min in the dark.
8. Start the reaction by prompt addition of 5 μL of either 350 units mL^{-1} of ADH or 5 units mL^{-1} of G6PDH as

appropriate. Mix the contents of the wells thoroughly and immediately place the plate in the microplate reader.

9. Run parallel controls with known amounts of each nucleotide (in the range 0.015–1.5 mM). Blanks are likewise included in the same plate by adding bicine · NaOH buffer instead of the sample. Blanks are particularly important in this experiment as they account for the amount of nucleotides bound to the enzymes used in the assay (although it is usually very low). For each standard, perform the procedure described above from **step 7** onward.
10. Monitor the formation of reduced MTT at 570 nm by recording the absorbance at 570 nm (A_{570}) and 30°C using a microplate reader.
11. Calculate the intracellular nucleotide concentration as follows. First, obtain the ΔA_{570} min^{-1} using the maximum linear rate for the controls, blank, and experimental samples for at least 10 min. Subtract the ΔA_{570} min^{-1} of the blank from the value obtained for the samples. Plot the ΔA_{570} min^{-1} of the standards against the nucleotide concentration in each of them to obtain a calibration curve in which the values for the samples are to be interpolated. Using the OD_{600} of the cultures, estimate the CDW concn. for each sample as detailed in Sect. 3.1.1. Obtain the nucleotide content by dividing the mole amount of each nucleotide by the CDW from which they were extracted.
12. Calculate the catabolic redox ratio as [NADH]/[NAD^+] and the anabolic redox ratio as [NADPH]/[$NADP^+$]. The total redox ratio is defined as [NADH] + [NADPH]/[NAD^+] + [$NADP^+$].

3.4 Oxidative Stress Test

For the determination of the sensitivity of the cells to oxidative stress, we propose a protocol in which oxidative stress conditions are imposed by adding the thiol-oxidizing agent DA [1,1′-azo-bis (*N*,*N*-dimethylformamide)] to the cultures from a concentrated solution. DA solutions are freshly prepared in DMSO. An appropriate volume of DMSO is added to control cultures, run in parallel.

1. Prepare *P. putida* cultures as described in the preceding sections.
2. Harvest 5–10-mL cell suspension aliquots at mid-exponential growth phase (at an OD_{600} of about 50% of the maximal value, OD_{600} = 0.5–0.6) by centrifugation (15 min, 4,000×*g*, 4°C). Wash the pellets twice with 10 mM $MgSO_4$.
3. Resuspend the cells in 10 mM $MgSO_4$ to an OD_{600} = 3.

4. Prepare a dose–response curve to determine the cells' sensitivity to diamide, by distributing M9 minimal medium supplemented with 20 mM glucose and increasing concentrations of DA (10–1,000 μg mL^{-1}) in 96-well microtiter plates. The DA concentration gradient is prepared by aliquoting the appropriate volume from the concentrated DA solution.
5. Inoculate each plate with the cell suspension adjusted to $OD_{600} = 3$, so that the initial OD_{600} is ~0.05.
6. Incubate the samples at 30°C using a microplate reader (with periodic agitation to prevent biomass sedimentation) and monitor the culture growth by turbidimetry (i.e., OD_{600}).
7. Determine the concentration of DA that produces a 50% and 100% inhibition of the bacterial growth. For this purpose, calculate the percentage of growth inhibition as $100 \times (\mu_{DA}/\mu_C)$, where μ_{DA} is the specific growth rate in the presence of any given concentration of DA and μ_C is the specific growth rate of the control culture (i.e., without any DA added). Specific growth rates are obtained as detailed in Sect. 3.1.1. Qualitatively correlate the redox ratios obtained in vitro as detailed in Sect. 3.3 (in particular, the anabolic redox ratio) to the tolerance of the cells to diamide.

4 Notes

1. There are several spectrophotometric- or HPLC-based analytical methods that can be used to determine the glucose content in liquid samples. Several commercial kits can also be used. Most of them are based on fluorescence (Amplex red glucose/glucose oxidase assay kit; Life Technologies) or colorimetry [glucose (GO) assay kit; Sigma-Aldrich Co.].
2. We have used an UltroSpec 3000 Pro UV–vis spectrophotometer (Biochrom Ltd., Cambridge, UK) for the determinations; however, the determination of kinetic parameters and enzymatic assays can be optimized in smaller volumes in a microplate reader [e.g., SpectraMax Plus 384 microplate reader or a Wallac 1420 VICTOR^2 multi-label counter and microplate reader (PerkinElmer Inc., Waltham, MA, USA)].
3. For labeling experiments, either 100% [1-^{13}C]-glucose or a mixture of 20% (w/w) [U-^{13}C]-glucose and 80% (w/w) natural glucose were used. [1-^{13}C]-Glucose was purchased from Cambridge Isotope Laboratories Inc. (Tewksbury, MA, USA) and [U-^{13}C]-glucose from Sigma-Aldrich Co.
4. A cell culture at metabolic steady state refers to the fact that there is no change neither in the growth rate nor in the uptake of carbon source over small periods of time. These growth

parameters usually remain constant in the range of few hours during exponential growth.

5. A SPB-1 capillary column [length = 30 m, internal diameter = 0.32 mm, film thickness = 0.25 μm, phase = poly(dimethyl siloxane), bonded] (Sigma-Aldrich Co., cat. # 24044) was used; however, other capillary nonpolar columns can also be employed.
6. GC–MS determinations were carried out in a Series 8000 gas chromatograph combined with a MD 800 mass spectrometer (Fisons Instruments PLC, Ipswich, UK). MD 800 mass spectrometer (Fisons Instruments) commonly comes equipped with Excalibur or Masslab software for data acquisition.
7. For flux calculations in MFA experiments, we recommend to use the MATLAB-based program Fiat Flux software [77].
8. A quartz cell of 1-cm width is the most commonly used for enzymatic assays involving NAD(P)($^+$/H). There are a number of plastic cells currently available that can be used in the UV region of the spectrum (e.g., BRAND UV cuvettes, BrandTech Scientific Inc., Essex, CT, USA, cat. # 759170).
9. A M9 salt mixture is prepared as a 10× concentrated solution and autoclaved. This 10× solution is obtained by mixing the following components (per liter of DI water): 128 g $Na_2HPO_4 \cdot 7H_2O$, 30 g KH_2PO_4, 5 g NaCl, and 10 g NH_4Cl [83]. Just prior to use, this concentrated solution is diluted with sterile DI water, added with $MgCl_2$, a trace element solution, and glucose as needed. $MgCl_2$ is added at 0.2 g L^{-1} from a filter-sterilized 2% (w/v) stock. A trace element solution [84] is added at 2.5 mL L^{-1}. Glucose is added at 20 mM from a filter-sterilized 20% (w/v) stock.
10. In this step, 0.20-μm nitrocellulose filters can also be used but the filtration process would be very slow. The procedure can also be done by centrifugation (15 min, 4,000×*g*, 4°C). After decanting the supernatant, the cells are subsequently washed with 0.9% (w/v) NaCl, transferred to pre-weighed Eppendorf tubes and dried at 105°C for 24 h to a constant weight.
11. To calculate the flux ratios, two independent experiments are required for each strain or condition: an experiment where the carbon source is 100% [1-^{13}C]-glucose and another experiment where the carbon source is a mixture of 20% (w/w) [U-^{13}C]-glucose and 80% (w/w) naturally labeled glucose.
12. The pellet may be stored at −20°C for several weeks.
13. Although a master network of biochemical reactions has been established and can be used for MFA experiments, the metabolic model should be adjusted as close as possible according to genomic information available for the microorganism under study.

14. Most enzyme assays are performed at 25°C; however, this protocol has been designed for *P. putida*, the optimum growth temperature of which is 30°C.
15. Depending on the activity of each enzyme, the time necessary to measure the change in A_{340} can vary widely. The time needed to get data in the linear portion of the absorbance versus time plot in these enzymatic assays usually varies between 5 and 10 min.
16. Depending on the culture conditions and the efficiency of protein extraction during the preparation of cell-free extracts, a dilution of the extract should also be tested in parallel to determine the best amount of total protein for each assay. Dilutions of the cell-free extract are prepared in the same buffer used for sonication.
17. In the case of activities represented by more than one enzyme (e.g., G6PDH), the cofactor specificity of the total activity is given.
18. We have found that keeping the time needed for cell harvesting and metabolic quenching to less than 3 min gives the best results in terms of recovery of pyridine nucleotides.
19. Store the neutralized samples at −20°C for no longer than 24 h.

References

1. Nicholson JK, Lindon JC (2008) Metabonomics. Nature 455:1054–1056
2. Joyce AR, Palsson BØ (2006) The model organism as a system: integrating "omics" data sets. Nat Rev Mol Cell Biol 7:198–210
3. Zhang W, Li F, Nie L (2010) Integrating multiple "omics" analysis for microbial biology: application and methodologies. Microbiology 156:287–301
4. Blankenburg M, Haberland L, Elvers H-D, Tannert C, Jandrig B (2009) High-throughput omics technologies: potential tools for the investigation of influences of EMF on biological systems. Curr Genomics 10:86–92
5. Gatherer D (2010) So what do we really mean when we say that systems biology is holistic? BMC Syst Biol 4:22
6. Kell DB (2004) Metabolomics and systems biology: making sense of the soup. Curr Opin Microbiol 7:296–307
7. Winter G, Krömer JO (2013) Fluxomics - connecting *omics* analysis and phenotypes. Environ Microbiol 15:1901–1916
8. Liu L, Agren R, Bordel S, Nielsen J (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Lett 584:2556–2564
9. Kitano H (2002) Systems biology: a brief overview. Science 295:1662–1664
10. Chavarría M, Kleijn RJ, Sauer U, Pflüger-Grau K, de Lorenzo V (2012) Regulatory tasks of the phosphoenolpyruvate-phosphotransferase system of *Pseudomonas putida* in central carbon metabolism. mBio 3, e00028-12
11. Fuhrer T, Fischer E, Sauer U (2005) Experimental identification and quantification of glucose metabolism in seven bacterial species. J Bacteriol 187:1581–1590
12. Sauer U, Eikmanns BJ (2005) The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria. FEMS Microbiol Rev 29:765–794
13. Dauner M, Bailey JE, Sauer U (2001) Metabolic flux analysis with a comprehensive isotopomer model in *Bacillus subtilis*. Biotechnol Bioeng 76:144–156
14. del Castillo T, Ramos JL, Rodríguez-Herva JJ, Fuhrer T, Sauer U, Duque E (2007) Convergent peripheral pathways catalyze initial glucose catabolism in *Pseudomonas putida*:

genomic and flux analysis. J Bacteriol 189:5142–5152

15. Berger A, Dohnt K, Tielen P, Jahn D, Becker J, Wittmann C (2014) Robustness and plasticity of metabolic pathway flux among uropathogenic isolates of *Pseudomonas aeruginosa*. PLoS One 9, e88368
16. Sauer U, Lasko DR, Fiaux J et al (1999) Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. J Bacteriol 181:6679–6688
17. Perrenoud A, Sauer U (2005) Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*. J Bacteriol 187:3171–3179
18. Stephanopoulos G (1999) Metabolic fluxes and metabolic engineering. Metab Eng 1:1–11
19. Blank LM, Ionidis G, Ebert BE, Bühler B, Schmid A (2008) Metabolic response of *Pseudomonas putida* during redox biocatalysis in the presence of a second octanol phase. FEBS J 275:5173–5190
20. Xiong W, Liu L, Wu C, Yang C, Wu Q (2010) ^{13}C-Tracer and gas chromatography–mass spectrometry analyses reveal metabolic flux distribution in the oleaginous microalga *Chlorella prototothecoides*. Plant Physiol 154:1001–1011
21. Shi H, Shiraishi M, Shimizu K (1997) Metabolic flux analysis for biosynthesis of poly(β-hydroxybutyric acid) in *Alcaligenes eutrophus* from various carbon sources. J Ferment Bioeng 84:579–587
22. Tyo KEJ, Fischer CR, Simeon F, Stephanopoulos G (2010) Analysis of polyhydroxybutyrate flux limitations by systematic genetic and metabolic perturbations. Metab Eng 12:187–195
23. Nanchen A, Fuhrer T, Sauer U (2007) Determination of metabolic flux ratios from ^{13}C-experiments and gas chromatography–mass spectrometry data: protocol and principles. Methods Mol Biol 358:177–197
24. Fischer E, Sauer U (2005) Large-scale *in vivo* flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. Nat Genet 37:636–640
25. Dauner M, Sauer U (2000) GC-MS analysis of amino acids rapidly provides rich information for isotopomer balancing. Biotechnol Prog 16:642–649
26. Sauer U (2006) Metabolic networks in motion: ^{13}C-based flux analysis. Mol Syst Biol 2:62
27. Fischer E, Zamboni N, Sauer U (2004) High-throughput metabolic flux analysis based on gas chromatography–mass spectrometry derived ^{13}C constraints. Anal Biochem 325:308–316
28. Fischer E, Sauer U (2003) Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. Eur J Biochem 270:880–891
29. Kleijn RJ, Buescher JM, Le Chat L, Jules M, Aymerich S, Sauer U (2010) Metabolic fluxes during strong carbon catabolite repression by malate in *Bacillus subtilis*. J Biol Chem 285:1587–1596
30. Meijnen JP, de Winde JH, Ruijssenaars HJ (2012) Metabolic and regulatory rearrangements underlying efficient D-xylose utilization in engineered *Pseudomonas putida* S12. J Biol Chem 287:14606–14614
31. Yang C, Hua Q, Shimizu K (2002) Metabolic flux analysis in *Synechocystis* using isotope distribution from ^{13}C-labeled glucose. Metab Eng 4:202–216
32. Nissen TL, Schulze U, Nielsen J, Villadsen J (1997) Flux distributions in anaerobic, glucose-limited continuous cultures of *Saccharomyces cerevisiae*. Microbiology 143:203–218
33. Christensen B, Nielsen J (1999) Isotopomer analysis using GC-MS. Metab Eng 1:282–290
34. Nikel PI, Zhu J, San KY, Méndez BS, Bennett GN (2009) Metabolic flux analysis of *Escherichia coli creB* and *arcA* mutants reveals shared control of carbon catabolism under microaerobic growth conditions. J Bacteriol 191:5538–5548
35. Emmerling M, Dauner M, Ponti A et al (2002) Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. J Bacteriol 184:152–164
36. Massou S, Nicolas C, Letisse F, Portais JC (2007) NMR-based fluxomics: quantitative 2D NMR methods for isotopomers analysis. Phytochemistry 68:2330–2340
37. Sekiyama Y, Kikuchi J (2007) Towards dynamic metabolic network measurements by multi-dimensional NMR-based fluxomics. Phytochemistry 68:2320–2329
38. Sauer U (2004) High-throughput phenomics: experimental methods for mapping fluxomes. Curr Opin Biotechnol 15:58–63
39. Nargund S, Joffe ME, Tran D, Tugarinov V, Sriram G (2013) Nuclear magnetic resonance methods for metabolic fluxomics. Methods Mol Biol 985:335–351
40. Chavarría M, Nikel PI, Pérez-Pantoja D, de Lorenzo V (2013) The Entner-Doudoroff pathway empowers *Pseudomonas putida* KT2440 with a high tolerance to oxidative stress. Environ Microbiol 15:1772–1785
41. Banga JR (2008) Optimization in computational systems biology. BMC Syst Biol 2:47

42. Fuhrer T, Sauer U (2009) Different biochemical mechanisms ensure network-wide balancing of reducing equivalents in microbial metabolism. J Bacteriol 191:2112–2121
43. Chen X, Li S, Liu L (2014) Engineering redox balance through cofactor systems. Trends Biotechnol 32:337–343
44. Singh R, Mailloux RJ, Puiseux-Dao S, Appanna VD (2007) Oxidative stress evokes a metabolic adaptation that favors increased NADPH synthesis and decreased NADH production in *Pseudomonas fluorescens*. J Bacteriol 189:6665–6675
45. 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
46. Berríos-Rivera SJ, Bennett GN, San KY (2002) The effect of increasing NADH availability on the redistribution of metabolic fluxes in *Escherichia coli* chemostat cultures. Metab Eng 4:230–237
47. Ruiz JA, de Almeida A, Godoy MS et al (2013) *Escherichia coli* redox mutants as microbial cell factories for the synthesis of reduced biochemicals. Comput Struct Biotechnol J 3, e201210019
48. Storz G, Imlay JA (1999) Oxidative stress. Curr Opin Microbiol 2:188–194
49. Cabiscol E, Tamarit J, Ros J (2000) Oxidative stress in bacteria and protein damage by reactive oxygen species. Int Microbiol 3:3–8
50. Carmel-Harel O, Storz G (2000) Roles of the glutathione- and thioredoxin-dependent reduction systems in the *Escherichia coli* and *Saccharomyces cerevisiae* responses to oxidative stress. Annu Rev Microbiol 54:439–461
51. Masip L, Veeravalli K, Georgiou G (2006) The many faces of glutathione in bacteria. Antioxid Redox Signal 8:753–762
52. Romano AH, Conway T (1996) Evolution of carbohydrate metabolic pathways. Res Microbiol 147:448–455
53. Downs DM (2006) Understanding microbial metabolism. Annu Rev Microbiol 60:533–559
54. Sudarsan S, Dethlefsen S, Blank LM, Siemann-Herzberg M, Schmid A (2014) The functional structure of central carbon metabolism in *Pseudomonas putida* KT2440. Appl Environ Microbiol 80:5292–5303
55. Conway T (1992) The Entner-Doudoroff pathway: history, physiology and molecular biology. FEMS Microbiol Rev 9:1–27
56. Poulsen BR, Nøhr J, Douthwaite S et al (2005) Increased NADPH concentration obtained by metabolic engineering of the pentose phosphate pathway in *Aspergillus niger*. FEBS J 272:1313–1325
57. Lee WH, Park JB, Park K, Kim MD, Seo JH (2007) Enhanced production of ε-caprolactone by overexpression of NADPH-regenerating glucose 6-phosphate dehydrogenase in recombinant *Escherichia coli* harboring cyclohexanone monooxygenase gene. Appl Microbiol Biotechnol 76:329–338
58. Marino D, González EM, Frendo P, Puppo A, Arrese-Igor C (2007) NADPH recycling systems in oxidative stressed pea nodules: a key role for the $NADP^+$-dependent isocitrate dehydrogenase. Planta 225:413–421
59. Rippa M, Giovannini PP, Barrett MP, Dallocchio F, Hanau S (1998) 6-Phosphogluconate dehydrogenase: the mechanism of action investigated by a comparison of the enzyme from different species. Biochim Biophys Acta 1429:83–92
60. Moritz B, Striegel K, De Graaf AA, Sahm H (2000) Kinetic properties of the glucose-6-phosphate and 6-phosphogluconate dehydrogenases from *Corynebacterium glutamicum* and their application for predicting pentose phosphate pathway flux *in vivo*. Eur J Biochem 267:3442–3452
61. Minard KI, McAlister-Henn L (2005) Sources of NADPH in yeast vary with carbon source. J Biol Chem 280:39890–39896
62. Miyagi H, Kawai S, Murata K (2009) Two sources of mitochondrial NADPH in the yeast *Saccharomyces cerevisiae*. J Biol Chem 284:7553–7560
63. Wynn JP, Ratledge C (1997) Malic enzyme is a major source of NADPH for lipid accumulation by *Aspergillus nidulans*. Microbiology 143:253–257
64. Ayala A, F-Lobato M, Machado A (1986) Malic enzyme levels are increased by the activation of NADPH-consuming pathways: detoxification processes. FEBS Lett 202:102–106
65. Blank LM, Lehmbeck F, Sauer U (2005) Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. FEMS Yeast Res 5:545–558
66. Remize F, Andrieu E, Dequin S (2000) Engineering of the pyruvate dehydrogenase bypass in *Saccharomyces cerevisiae*: role of the cytosolic Mg^{2+} and mitochondrial K^+ acetaldehyde dehydrogenases Ald6p and Ald4p in acetate formation during alcoholic fermentation. Appl Environ Microbiol 66:3151–3159
67. Nikel PI, Kim J, de Lorenzo V (2014) Metabolic and regulatory rearrangements underlying glycerol metabolism in *Pseudomonas putida* KT2440. Environ Microbiol 16:239–254

68. Sauer U, Canonaco F, Heri S, Perrenoud A, Fischer E (2004) The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of *Escherichia coli*. J Biol Chem 279:6613–6619
69. Rühl M, Le Coq D, Aymerich S, Sauer U (2012) ^{13}C-Flux analysis reveals NADPH-balancing transhydrogenation cycles in stationary phase of nitrogen-starving *Bacillus subtilis*. J Biol Chem 287:27959–27970
70. Singh R, Lemire J, Mailloux RJ, Appanna VD (2008) A novel strategy involved anti-oxidative defense: the conversion of NADH into NADPH by a metabolic network. PLoS One 3, e2682
71. Sprenger GA (1995) Genetics of pentose-phosphate pathway enzymes of *Escherichia coli* K-12. Arch Microbiol 164:324–330
72. Dean AM, Golding GB (1997) Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. Proc Natl Acad Sci U S A 94:3104–3109
73. Eyzaguirre J, Cornwell E, Borie G, Ramirez B (1973) Two malic enzymes in *Pseudomonas aeruginosa*. J Bacteriol 116:215–221
74. Nelson KE, Weinel C, Paulsen IT et al (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. Environ Microbiol 4:799–808
75. Nikel PI, Martínez-García E, de Lorenzo V (2014) Biotechnological domestication of pseudomonads using synthetic biology. Nat Rev Microbiol 12:368–379
76. Zamboni N, Fendt S-M, Rühl M, Sauer U (2009) ^{13}C-based metabolic flux analysis. Nat Protoc 4:878–892
77. Zamboni N, Fischer E, Sauer U (2005) FiatFlux - a software for metabolic flux analysis from ^{13}C-glucose experiments. BMC Bioinformatics 6:209
78. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72:248–254
79. Bernofsky C, Swan M (1973) An improved cycling assay for nicotinamide adenine dinucleotide. Anal Biochem 53:452–458
80. Nikel PI, Pettinari MJ, Ramirez MC, Galvagno MA, Méndez BS (2008) *Escherichia coli arcA* mutants: metabolic profile characterization of microaerobic cultures using glycerol as a carbon source. J Mol Microbiol Biotechnol 15:48–54
81. Nikel PI, Pettinari MJ, Galvagno MA, Méndez BS (2010) Metabolic selective pressure stabilizes plasmids carrying biosynthetic genes for reduced biochemicals in *Escherichia coli* redox mutants. Appl Microbiol Biotechnol 88:563–573
82. Leonardo MR, Dailly Y, Clark DP (1996) Role of NAD in regulating the *adhE* gene of *Escherichia coli*. J Bacteriol 178:6013–6018
83. Miller JH (1972) Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor
84. Nikel PI, de Lorenzo V (2013) Engineering an anaerobic metabolic regime in *Pseudomonas putida* KT2440 for the anoxic biodegradation of 1,3-dichloroprop-1-ene. Metab Eng 15:98–112

Design of Orthogonal Pairs for Protein Translation: Selection Systems for Genetically Encoding Noncanonical Amino Acids in *E. coli*

Jelena Jaric and Nediljko Budisa

Abstract

The expansion of the genetic code is gradually becoming a core discipline in synthetic biology. Residue-specific incorporation of noncanonical amino acids (ncAAs) into proteins allows facile alteration and enhancement of protein properties. There are two distinct in vivo approaches available for their cotranslational incorporation. For isostructural noncanonical amino acids, residue-specific replacement of canonical amino acids is performed with the supplementation-based incorporation method (SPI) using auxotrophic host strains. On the other hand, orthogonal ncAAs are incorporated into the proteins site specifically in response to stop or quadruplet codons (stop codon suppression (SCS)) using orthogonal aminoacyl-tRNA synthetase/tRNA pairs (o-pair). Frequently used o-pair is based on the tyrosyl-tRNA synthetase from *Methanocaldococcus jannaschii* (MjTyrRS). To evolve a new orthogonal aminoacyl-tRNA synthetase (aaRS), which recognizes exclusively the noncanonical amino acid, the most straightforward solution is to produce a library of MjTyrRS mutants, containing randomized residues in the amino acid-binding site, on the basis of available crystal structure. The library is transformed into *Escherichia coli* and three rounds of positive and negative selection are performed in order to select for desired MjTyrRS variant which uniquely charges the tRNA with the ncAA of interest. Here, we provide a protocol with detailed description how to perform positive and negative selection with chloramphenicol acetyltransferase and barnase, respectively.

Keywords: Amber suppressor *Methanocaldococcus jannaschii* $tRNA^{Tyr}_{CUA}$opt, *Methanocaldococcus jannaschii* tyrosyl-tRNA synthetase, Noncanonical amino acid, Orthogonal pair, Positive and negative selection, Stop codon suppression approach

1 Introduction

To generate natural proteins, only 20 canonical amino acids (cAAs), encoded by the 61 sense codons, are used. However, due to the limited range of functions performed by proteins, they often require amino acid side chains with increased chemical functionality. The main source of chemical diversity in the majority of mature

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 71–82,
DOI 10.1007/8623_2015_105, © Springer-Verlag Berlin Heidelberg 2015, Published online: 19 June 2015

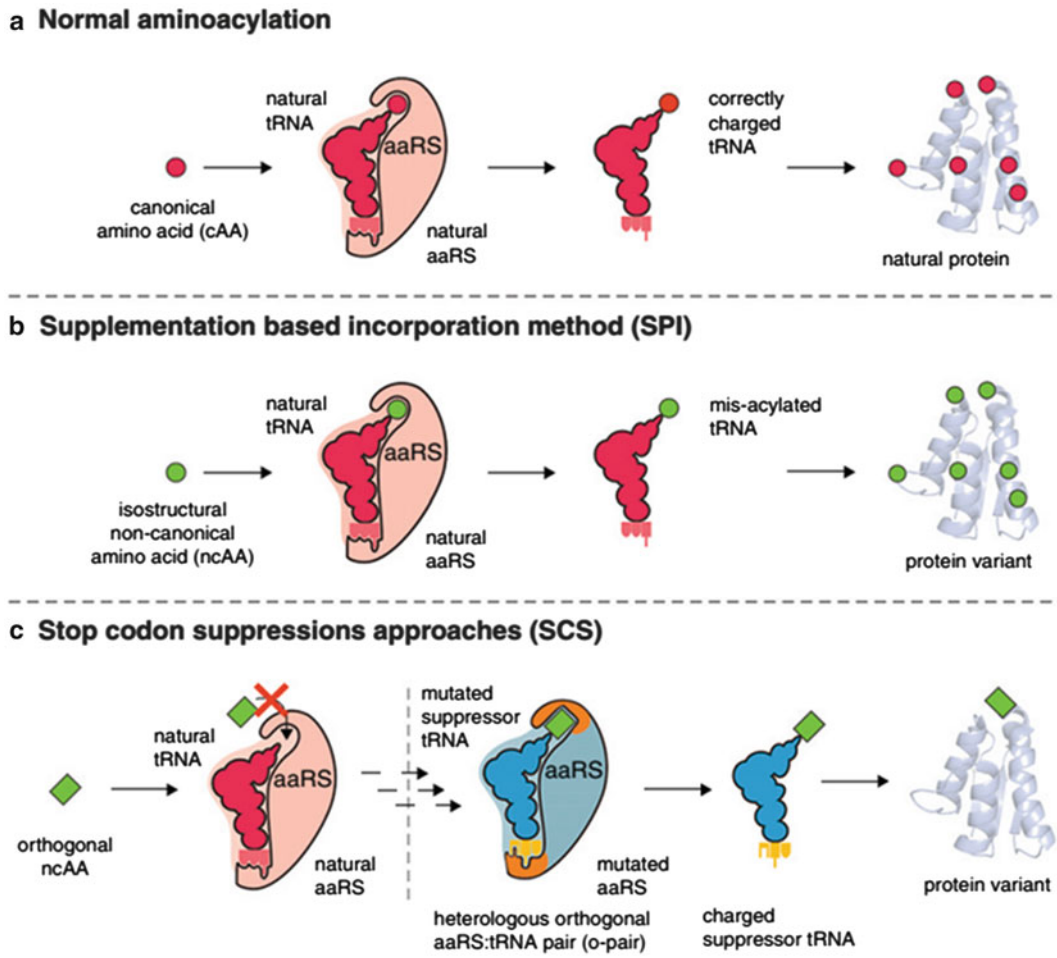


Fig. 1 Aminoacylation with canonical and noncanonical amino acids for protein translation. (**a**) In normal aminoacylation reaction aminoacyl-tRNA synthetase charges tRNAs with the corresponding cognate amino acid. (**b**) Supplementation-based incorporation method (SPI). (**c**) Stop codon suppression (SCS) approaches. See introduction part for more detailed description (Figure reproduced from Hoesl M.G. and Budisa N. 2012 with permission from Elsevier [2])

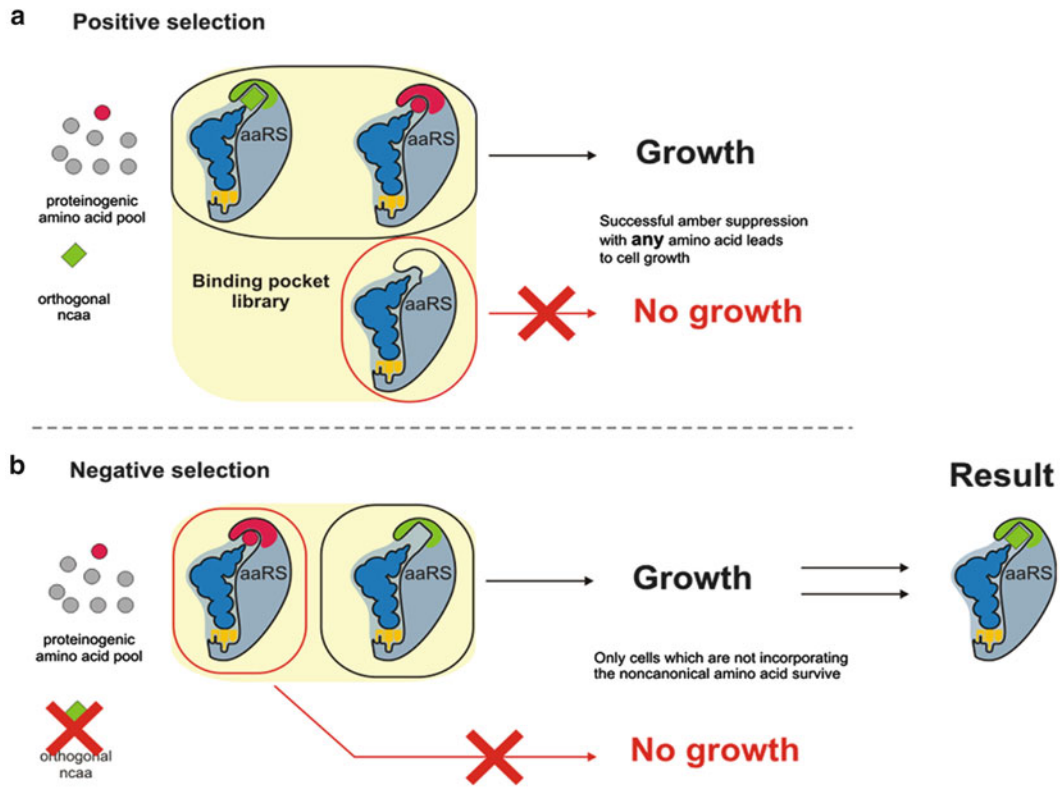
proteins and peptides is posttranslational modifications (PTMs). These complex processes, performed by enzymes and enzyme assemblies, are separated from translation concerning both time and space. Thus, the engineering of cells with expanded genetic codes that include noncanonical amino acids (ncAAs) allows design of proteins with enhanced and novel characteristics and activities [1].

Currently, there are two different approaches for the incorporation of noncanonical amino acids into the proteins (Fig. 1). The first approach, so-called supplementation-based incorporation (SPI), relies on the natural substrate tolerance of the endogenous host aminoacyl-tRNA synthetase (aaRS) by using auxotrophic host

strain (Fig. 1b). In this way, noncanonical amino acids which are isostructural to their canonical counterparts can be incorporated in a target protein via sense-codon reassignment [3]. Stop codon suppression (SCS) methodology, the second approach, uses a heterologous orthogonal aaRS:tRNA pair (o-pair) to incorporate an orthogonal amino acid in response to a stop codon site specifically (Fig. 1c). Orthogonality is a crucial condition and is defined by a lack of cross-reactivity between the o-pair (including the ncAA) and the endogenous host aminoacyl-tRNA synthetases, amino acid, and tRNAs [4].

Since *Methanocaldococcus jannaschii* tyrosyl-tRNA synthetase (MjTyrRS) and its cognate $tRNA^{Tyr}$ ($MjtRNA^{Tyr}$) are almost a natural orthogonal pair in *E. coli* [5], the MjTyrRS is often used for evolving an aaRS, which will exclusively charge tRNA with desired ncAA, by introducing mutations in the active site of MjTyrRS. Normally, up to five amino acids are randomized into all 20 amino acids and several are placed into fixed mutations based on the available crystal structure [6] and rational design. Additionally, since the incorporation of ncAA is in response to a stop codon, $MjtRNA^{Tyr}$ has to be mutated into amber suppressor tRNA by changing the tRNA anticodon to CUA. In 2009 Guo et al. performed directed evolution experiments that focused on the T-stem of amber suppressor $MjtRNA^{Tyr}_{CUA}$ and identified a modified optimized suppressor ($MjtRNA^{Tyr}_{CUA}$opt) that increased unnatural amino acid incorporation efficiency with several aaRS [7].

Once the MjTyrRS library is produced, a few rounds (usually three) of positive and negative selection have to be performed in order to select for the MjTyrRS variant which uniquely recognizes the ncAA of interest (Fig. 2). MjTyrRS library is transformed into *E. coli* cells that express $MjtRNA^{Tyr}_{CUA}$opt and a gene encoding chloramphenicol acetyltransferase (CAT) with two amber stop codons at a permissive sites (Fig. 3a). In this step cells are grown in the presence of all canonical amino acids, noncanonical amino acids, and chloramphenicol so that only cells with the aaRS mutants capable of aminoacylating $MjtRNA^{Tyr}_{CUA}$opt with the ncAA or any endogenous amino acids live since suppression of amber stop codons in *cat* gene gives cells the resistance to chloramphenicol (Cm). Surviving mutants are then transformed into *E. coli* cells that express $MjtRNA^{Tyr}_{CUA}$opt and the toxic barnase gene with two amber stop codons at permissive sites (Fig. 3b). This is negative selection step: cells are grown in the presence of all canonical amino acids, but in the absence of noncanonical amino acid. In this way, cells which contain aaRS mutants which still charge suppressor tRNA with any of 20 canonical amino acids die since the suppression of amber stop codons in the gene for barnase results with the expression of toxic protein. This leaves only aaRS variant that aminoacylates $MjtRNA^{Tyr}_{CUA}$opt with the noncanonical amino acid of interest [4].



c Substrates

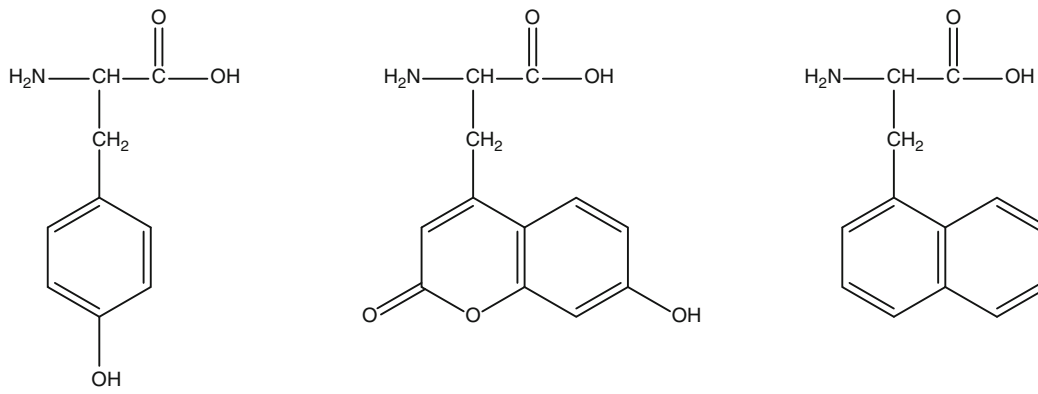


Fig. 2 Principle of positive and negative selection. (**a**) In positive selection only aaRS mutants which charge suppressor tRNA with any of 20 canonical amino acids or noncanonical amino acid lead to cell growth. In this step mutants whose amino acid-binding pocket is inactive are discriminated. (**b**) In negative selection only cells which are not incorporating the noncanonical amino acid survive since ncAA is not provided in the growth media. All other aaRS mutants which still charge suppressor tRNA with any of 20 canonical amino acids enable translation of barnase which leads to cell death. Graphic kindly provided by Dr. Michael Hösl. (**c**) Substrates to be tested are tyrosine which is natural or canonical amino acid whereas fluorescent probes coumarin alanine and 1-naphthylalanine are noncanonical counterparts

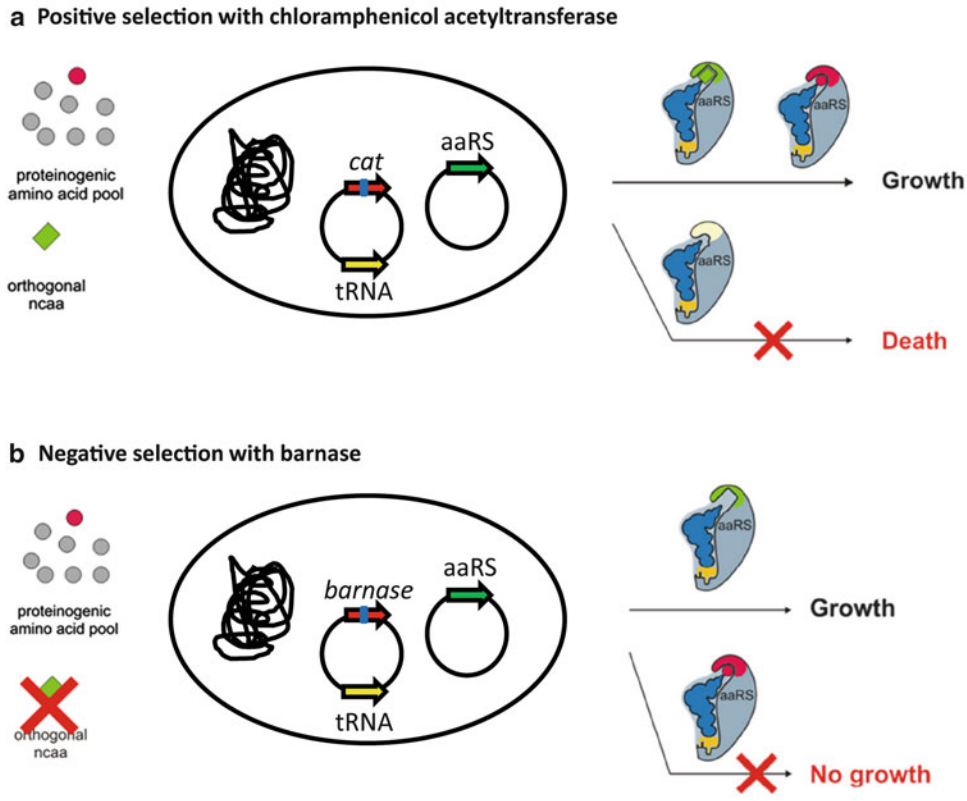


Fig. 3 CAT/barnase-based selection system. (**a**) Positive selection with chloramphenicol acetyltransferase. Cells grow in the presence of all canonical amino acids (*red* and *gray circles*), noncanonical amino acid (*green square*), and chloramphenicol so that only cells with the aaRS mutants capable of aminoacylating MjtRNA$^{Tyr}_{CUA}$opt with the ncAA or any endogenous amino acid live. (**b**) Negative selection with barnase. Cells grow in the absence of noncanonical amino acid (*green square*). In this way, cells which contain aaRS mutants which still charge suppressor tRNA with any of 20 canonical amino acids rescue barnase activity and subsequently die. Therefore only cells with aaRS variant that exclusively aminoacylates MjtRNA$^{Tyr}_{CUA}$opt with the noncanonical amino acid of interest survive negative selection step

In this chapter we describe how to perform positive and negative selection with chloramphenicol acetyltransferase and barnase, respectively, with the already selected MjTyrRS variant in order to prove its substrate specificity and functionality. We exemplify here the library design with fluorescent ncAAs coumarin alanine and 1-naphthylalanine (Fig. 2). Although their orthogonal pairs have been reported [4], there is still a great need for the catalytic improvements.

From a practical standpoint, it is important to emphasize that reassignment of stop codons is not toxic to cells, as a whole procedure is in fact an extension of the recombinant DNA technology. Namely, the cells are grown to mid-log phase, and after enough cell mass is accumulated, plasmid encoded gene with in-frame stop

codon is almost exclusively expressed. The widely used promoters (e.g., T7, T5, arabinose, etc.) do not support cellular growth of the bacterial host after induction of the protein synthesis at all, i.e., the whole cellular machinery is in the function of target protein production. Bacterium *Escherichia coli* is commonly used as host cell, although other microorganisms (e.g., yeast) and even higher eukaryotic cells (from insects, mammals) were used as well [4].

2 Materials

2.1 Preparation of 24-Well Plate for Positive Selection

1. New minimal media without tyrosine (NMM (−Tyr)) with agar (*see* Subheading 2.6).
2. Antibiotics: ampicillin, kanamycin, and chloramphenicol (*see* Subheading 2.5).
3. 24-well plate (Techno Plastic Products; TPP (http://www.tpp.ch)).
4. Noncanonical amino acid (*see* Subheading 2.5) (*see* **Note 1**).

2.2 Transformation of Chemically Competent E. coli DH10b Cells

1. Vector: pBU181GK_MjTyrRSmutant (Fig. 4c) (*see* **Note 2**).
2. Bacteria strain: chemically competent *E. coli* DH10b strain (Life Technologies (http://www.lifetechnologies.com)) which already contains plasmid pPAB26′_cat(Q98TAG, D181TAG) MjtRNA$^{Tyr}_{CUA}$opt or pNB26′2_barnase(Q2TAG, D44TAG) MjtRNA$^{Tyr}_{CUA}$opt, respectively (Fig. 4a, b).
3. LB media (*see* Subheading 2.6).

2.3 Cell Growth and Positive Selection with Chloramphenicol Acetyltransferase

1. LB media (*see* Subheading 2.6).
2. Antibiotics: ampicillin and kanamycin (*see* Subheading 2.5).
3. Bacteria strain: *E. coli* DH10b strain (Life Technologies (http://www.lifetechnologies.com)) transformed with two

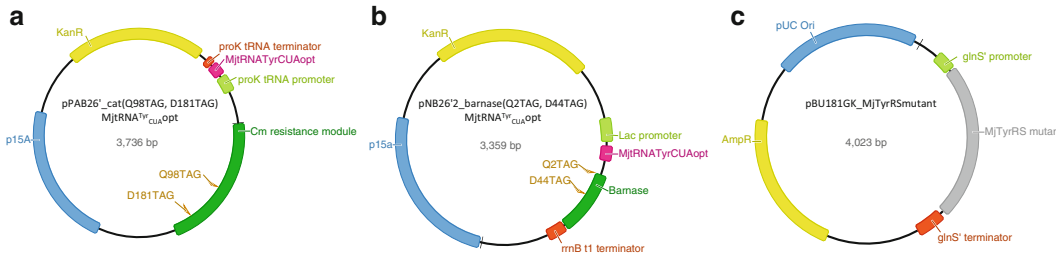


Fig. 4 Maps of plasmids used in our laboratory. (**a**) pPAB26′_cat(Q98TAG, D181TAG) MjtRNA$^{Tyr}_{CUA}$opt is kanamycin-resistance plasmid carrying genes for CAT with two in-frame amber stop codons (Q98TAG and D181TAG) and MjtRNA$^{Tyr}_{CUA}$opt. (**b**) pNB26′2_barnase(Q2TAG, D44TAG) MjtRNA$^{Tyr}_{CUA}$opt is kanamycin-resistance plasmid carrying genes for barnase with two in-frame amber stop codons (Q2TAG and D44TAG) and MjtRNA$^{Tyr}_{CUA}$opt. (**c**) pBU181GK_MjTyrRSmutant is ampicillin-resistance plasmid carrying gene for MjTyrRS mutant

plasmids: pPAB26′_cat(Q98TAG, D181TAG) MjtRNATyr-CUAopt and pBU181GK_MjTyrRSmutant (Fig. 4a, c).

4. NMM (−Tyr) (*see* Subheading 2.6).
5. Previously prepared 24-well plate.

2.4 Negative Selection with Barnase

1. Plates: LBAmpKan with 2 mM ncAA and LBAmpKan.
2. Vector: pBU181GK_MjTyrRSmutant (Fig. 4c) (*see* **Note 2**).
3. Bacteria strain: chemically competent *E. coli* DH10b strain (Life technologies (http://www.lifetechnologies.com)) which already contains plasmid pNB26′2_barnase(Q2TAG, D44TAG) MjtRNA$^{Tyr}_{CUA}$opt (Fig. 4b).
4. LB media (*see* Subheading 2.6).

2.5 General Buffers and Reagents

1. Ampicillin (Roth (http://www.carlroth.com)): 100 mg/ml in water. Store at −20°C.
2. Kanamycin (Roth (http://www.carlroth.com)): 50 mg/ml in water. Store at −20°C.
3. Chloramphenicol (Roth (http://www.carlroth.com)): 37 mg/ml in 100% ethanol. Store at −20°C.
4. Noncanonical amino acid: if possible, prepare 40 mM stock and store according to manufacturer's instructions.

2.6 Bacteria Growth Media

1. LB: 10 g tryptone/peptone, 5 g yeast extract, and 5 g NaCl/L water.
2. New minimal media (NMM): 7.5 mM $(NH_4)_2SO_4$, 8.5 mM NaCl, 22 mM KH_2PO_4, 50 mM K_2HPO_4, 1 mM $MgSO_4$, 20 mM D-glucose, 50 mg/L all amino acids (*see* **Note 3**), 1 μg/mL Ca^{2+}, 1 μg/mL Fe^{2+}, 0.01 μg/mL trace elements (Cu^{2+}, Zn^{2+}, Mn^{2+}, $MoOH^{2+}$), 10 μg/mL thiamine, 10 μg/mL biotin, 100 μg/mL ampicillin, 50 μg/mL kanamycin, 0.5–2 mM noncanonical amino acid (*see* **Note 4**)

To prepare solid LB media, agar at the final concentration of 15 g/L was added to the solution. Following the autoclaving, the media was supplemented with needed antibiotics. The final concentrations of the antibiotics used in this study were as follows: ampicillin 100 μg/mL, kanamycin 50 μg/mL, and chloramphenicol 37 μg/mL. To prepare solid NMM (−Tyr), freshly autoclaved (still liquid) 2 × agar (30 g/L) was gently mixed with the equal volume of 2 × NMM (−Tyr); the final concentration of NMM components in 2 × NMM (−Tyr) is as twice as mentioned above (*see* **Note 5**).

3 Methods

3.1 Positive Selection with Chloramphenicol Acetyltransferase

3.1.1 Preparation of 24-Well Plate

1. Pipette in each well 100 μL autoclaved sterile MilliQ water.
2. Add 100 μL 40 mM noncanonical amino acid in the wells from A1 to B6 so that the final concentration of ncAA in the final volume of 2 mL is 2 mM (Fig. 5).
3. Instead of ncAA add the same volume (100 μL) of autoclaved sterile MilliQ water in the wells from C1 to D6.
4. Pipette 100 μL Cm of appropriate concentration in the wells from A1 to B6 and from C1 to D6, respectively, so that the concentration of Cm rises in the following way: 0, 5, 10, 15, 25, 37, 50, 60, 75, 100, 150, and 200 μg/mL (Fig. 5).
5. In the end add 1.7 mL NMM (−Tyr) with agar and premixed ampicillin and kanamycin, mix carefully to avoid air bubbles in the media, and let the 24-well plate stand on the room temperature (RT) to cool down.

3.1.2 Transformation of Chemically Competent E. coli DH10b Cells

In order to perform positive selection test, chemically competent *E. coli* DH10b cells which already contain plasmid pPAB26′_cat (Q98TAG, D181TAG) MjtRNA$^{Tyr}_{CUA}$opt were transformed with pBU181GK_MjTyrRSmutant in the following way:

1. Add 100–500 ng of plasmid DNA pBU181GK_MjTyrRSmutant into 50 μL *E. coli* DH10b cells.
2. Incubate on ice 30 min.
3. Incubate at 42°C 2 min.
4. Add immediately 950 μL LB media.
5. Incubate at 37°C 60 min.

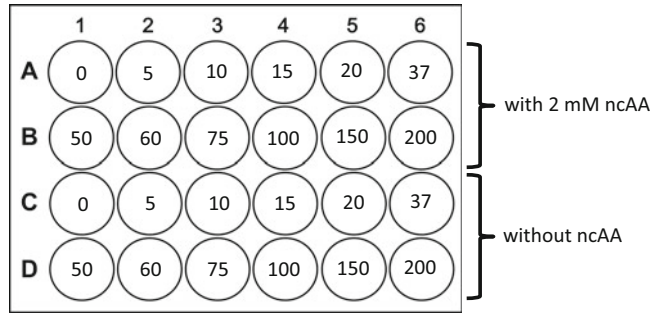


Fig. 5 Schematic representation of 24-well plate with antibiotic concentration gradient. Wells in the lines A (A1–A6) and B (B1–B6) are supplemented with 2 mM noncanonical amino acid, while wells in the lines C (C1–C6) and D (D1–D6) do not contain noncanonical amino acid. *Different numbers in the circles* designate the chloramphenicol concentration in μg/mL

3.1.3 Cell Growth and Test on the 24-Well Positive Selection Plate

1. Inoculate 9 mL of LB which contains 100 μg/mL ampicillin and 50 μg/mL kanamycin (LBAmpKan) with 1 mL of freshly transformed *E. coli* DH10b cells from the previous step.
2. Grow the cells shaking (200 rpm) overnight at 37°C to an $OD_{600nm\ (optical\ density)} = 0.6$.
3. Spin down 1 mL of cells by centrifugation at maximal speed (13,400 rpm) for 2 min.
4. Resuspend the cells gently in 1 mL of NMM (−Tyr) by pipetting up and down and spin down the cells by centrifugation at maximal speed (13,400 rpm) for 2 min. Remove the supernatant.
5. Repeat step 4.
6. Resuspend the cells in 1 mL of NMM (−Tyr).
7. Plate 10 μL of cells on each well on the 24-well plate.
8. Incubate at 37°C overnight (*see* **Note 6**) (Fig. 6a).

Figure 6a represents 24-well plate after 1-day incubation at 37°C. Cells grow up to 37 μg/mL Cm in the presence of ncAA (well A6), while in the absence of ncAA, cells grow only on the NMM (−Tyr) media without chloramphenicol (0 μg/mL Cm; well C1), as expected. After 2-day incubation at 37°C, there is cell growth up to 150 μg/mL Cm and 2 mM ncAA (well B5) and up to 15 μg/mL Cm when ncAA is omitted (well C4) (Fig. 6b).

3.2 Negative Selection with Barnase

1. Prepare two different type of plates: (1) LBAmpKan with 2 mM ncAA and (2) LBAmpKan.
2. Transform chemically competent *E. coli* DH10b cells which already contain plasmid pNB26′2_barnase(Q2TAG, D44TAG) $MjtRNA^{Tyr}_{CUA}$opt with pBU181GK_MjTyrRSmutant as described under "Transformation of chemically competent *E. coli* DH10b cells" (*see* Subheading 3.1.2).
3. After 1 h incubation at 37°C (*see* Subheading 3.1.2), plate 20, 200, and 780 μL of cells on LBAmpKan with/without ncAA plates (*see* **Note 7**) (Fig. 7).
4. The next day, streak several colonies (e.g., 16 colonies) from the LBAmpKan plate to a fresh plate of the same media composition. Incubate the plate at 37°C overnight (Fig. 8a).
5. To confirm once more that streaked colonies do not grow on LBAmpKan with noncanonical amino acid, restreak them on a such plate. Incubate the plate at 37°C overnight (Fig. 8b).

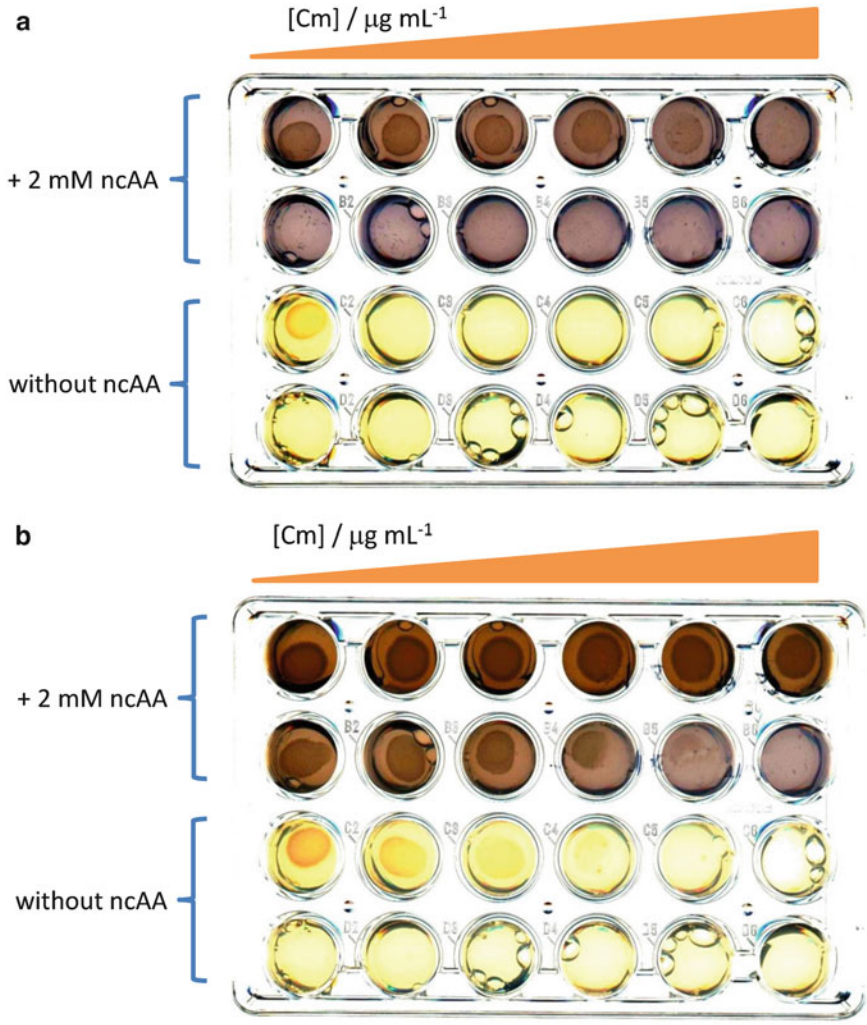


Fig. 6 Positive selection 24-well plates after (**a**) 1-day and (**b**) 2-day incubation. *E. coli* DH10b cells were transformed with pPAB26′_cat(Q98TAG, D181TAG) MjtRNA$^{Tyr}_{CUA}$opt and pBU181GK_MjTyrRSmutant. *Rows A and B* (A1–A6 and B1–B6) contain 2 mM ncAA, while *rows C and D* (C1–C6 and D1–D6) do not contain ncAA. Chloramphenicol concentration varies like shown in schematic representation in Fig. 5. Cells were left to grow on NMM (−Tyr) Amp Kan at 37°C for 1 or 2 days, respectively

4 Notes

1. We prefer not to define the exact name of noncanonical amino acid since its incorporation into the proteins and respective new orthogonal aminoacyl-tRNA synthetase are not published yet. So, we keep the name of used noncanonical amino acid as "noncanonical amino acid or ncAA" throughout the text.
2. The vectors should be purified by mini or midi prep kit.

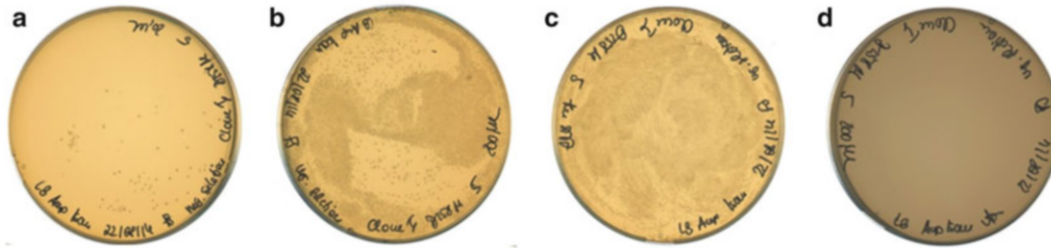


Fig. 7 Negative selection plates in the (**a**) (**b**) (**c**) absence and (**d**) presence of noncanonical amino acid. Different amounts of cells from the culture with $OD_{600} \cong 1$ were plated: (**a**) 20 μL, (**b**) 200 μL, and (**c**) 780 μL on the LBAmpKan plates. Cells grew at 37°C overnight. Due to the lack of ncAA, expression of toxic barnase does not occur and cells live. (**d**) Only LBAmpKan with 2 mM ncAA plate with 200 μL of cells is shown. Cells grew at 37°C overnight. Due to the presence of ncAA, toxic barnase is expressed and cells are not growing. *E. coli* DH10b cells were transformed with pNB26′2_barnase(Q2TAG, D44TAG) $MjtRNA^{Tyr}_{CUA}$opt and pBU181GK_MjTyrRSmutant

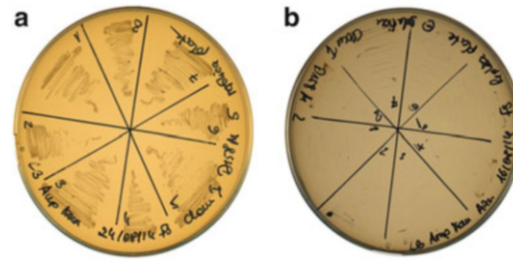


Fig. 8 Control experiment on negative selection plates in the (**a**) absence and (**b**) presence of noncanonical amino acid, respectively. (**a**) Eight colonies from the negative selection LBAmpKan plates (Fig. 7a, b, c) were streaked on the LBAmpKan plate and incubated at 37°C overnight. (**b**) The same eight colonies were restreaked on the LBAmpKan with 2 mM ncAA and incubated at 37°C overnight. As expected, cellular growth is completely inhibited in the presence of ncAA

3. We used NMM without tyrosine (NMM (−Tyr)). When evolving a new aaRS from MjTyrRS or testing the existing one, one can omit tyrosine when preparing NMM for positive selection.
4. Depending on the commercial availability and price, one can also use higher concentrations of noncanonical amino acid (e.g., 5 mM).
5. Mix carefully 2 × NMM (−Tyr) and 2 × agar to avoid air bubbles in liquid media and afterward in solid media in the Petri dish or 24-well plate.
6. It's recommended to incubate the plates up to 3 days to follow the growth difference among wells with and without ncAA and wells with different Cm concentration.

7. After plating 20 and 200 μL of cells from the total volume of 1 mL, we advise to spin down the remaining 780 μL of cells by centrifugation at maximal speed (13,400 rpm) for 1–2 min. Resuspend the cell pellet in 100–200 μL of LB and plate.

References

1. Budisa N (2013) Expanded genetic code for the engineering of ribosomally synthetized and post-translationally modified peptide natural products (RiPPs). Curr Opin Biotechnol 24:591–598
2. Hoesl MG, Budisa N (2012) Recent advances in genetic code engineering in *Escherichia coli*. Curr Opin Biotechnol 23:751–757
3. Link AJ, Mock ML, Tirrell DA (2003) Non-canonical amino acids in protein engineering. Curr Opin Biotechnol 14:603–609
4. Liu CC, Schultz PG (2010) Adding new chemistries to the genetic code. Annu Rev Biochem 79:413–444
5. Wang L, Brock A, Herberich B, Schultz PG (2001) Expanding the genetic code of *Escherichia coli*. Science 292:498–500
6. Zhang Y, Wang L, Schultz PG, Wilson IA (2005) Crystal structures of apo wild-type *M. jannaschii* tyrosyl-tRNA synthetase (TyrRS) and an engineered TyrRS specific for O-methyl-L-tyrosine. Protein Sci 14:1340–1349
7. Guo J, Melancon CE III, Lee HS, Groff D, Schultz PG (2009) Evolution of amber suppressor tRNAs for efficient bacterial production of proteins containing nonnatural amino acids. Angew Chem Int Ed 48:9148–9151

Phenome-ing Microbes

Klaus Hornischer and Susanne Häussler

Abstract

One of the burning questions in bacterial genomics is how the phenotype of a bacterial strain correlates to its genotype. Some phenotypes of a given organism's isolate arise through simple sequence variations like single nucleotide polymorphisms (SNP) or small insertions/deletions (InDel). For some phenotypes, however, the underlying mechanism cannot be explained by simple genomic differences; rather, most of them are the result of more complex sequence variations. Insight into complex phenotypes such as bacterial pathogenicity, or resistance traits and their molecular background, require comprehensive data obtained in large-scale projects and involve statistical methods. With the increasing usage of next-generation sequencing (NGS) and other "-omics" techniques in molecular biology, projects are now feasible which provide such a data foundation. Big data, however, not only offers new opportunities but also requires extensive data management systems. A coupled system of a relational database, web interface and statistical methods provides substantial support for phenotype-genotype correlation studies aimed to unravel molecular mechanisms underlying complex phenotypes and designed for biomarker identification.

Keywords Association study, Biomarker identification, Genotype-phenotype correlation

1 Introduction

One of the major questions posed and addressed, especially in large-scale sequencing projects, is how the genotype of a bacterial strain correlates to its phenotypes. Some phenotypes of strains or isolates of a given organism are explained by one small difference in the nucleotide sequence (SNP or InDel) in comparison to a reference strain, which lacks the respective trait. For most phenotypic differences, however, the underlying mechanism cannot be explained by such simple genomic variation. Most of the phenotypic traits result from the combination of several sequence differences.

The identification of a set of genetic determinants that are most likely involved in the formation of a phenotype is not a simple task and strictly relies on statistical methods. Due to high sequencing costs in past years, the amount of data required for robust statistics was frequently not available. With the recent introduction of new sequencing and other "-omics" technologies [1, 2], a wealth of new

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 83–96,
DOI 10.1007/8623_2015_178, © Springer-Verlag Berlin Heidelberg 2015, Published online: 15 December 2015

possibilities has been introduced into molecular biology. Experimental designs, which were not realistic in the past, due to high costs or lengthy project times, are now feasible [3, 4].

One potential pitfall of association studies is the creation of false positive associations due to small sample size or a lack of comparability of the two groups to be compared. Only studies that are based on an appropriately high number of isolates displaying the phenotype under investigation, compared to a properly selected group of control variants lacking the phenotype, have a good chance to produce meaningful results. In an ideal experimental design, the two isolate groups upon which the study is based would be identical in terms of ancestry, environmental conditions and phenotypes, with the exception of the phenotype under consideration. This is obviously difficult to achieve. In such a situation, a large sample size and a collection of samples from very heterogeneous sources minimises the probability of spurious associations as unlinked genetic differences level out due to the big sample size, and associations which are statistically detected have a higher probability to be true positives.

Recording of a plethora of biological features by the use of "-omics" technologies can substantially facilitate the differentiation of phenotypes ("stratification"). The term stratification is used in personalised medicine for the identification of patient groups which share "biological" characteristics, through biochemical, molecular and imaging diagnostic test methods [5]. In the context of molecular bacteriology, this means that "-omics" data on the bacterial genome of a given species and its derivatives such as RNA, protein and metabolites is used for the classification and characterisation of bacterial phenotypes.

The aim of new approaches for genotype-phenotype correlation studies is to provide extensive insights into the underlying molecular mechanisms of even very complex bacterial phenotypes. Furthermore, the subsequent identification of biomarkers that serve stratification purposes, for instance susceptible versus non-susceptible or pathogenic versus non-pathogenic isolates, is expected to change and significantly advance routine medical microbiology diagnostic procedures with respect to predictive power and due to a lower "turnaround time".

The data sets resulting from genome sequencing, protein expression profiling, metabolomics and RNA-seq [6] experiments are very large. Although for a small number of experiments values can still be stored and handled in spreadsheets (e.g. Microsoft Excel or OpenOffice Calc), for larger data collections the storage and analysis is much more complicated or only feasible in a relational database system. The structure of relational databases facilitates data filtering, which has to be applied to produce meaningful subsets of the data. The extracted data sets are then subjected to statistical analysis, with the aim to identify statistically relevant associations between sequence variations and a particular

phenotype. The analysis results form the start point for the development of hypotheses explaining the molecular mechanisms of phenotypic traits, which have to be subsequently validated in wet lab experiments. The accumulated knowledge can be exploited for the identification of potential biomarkers that serve stratification purposes [7].

2 Materials and Methods

2.1 Database Creation

Although it is feasible to handle data resulting from a small number of transcriptomics, proteomics and/or metabolomics experiments in flat files like spreadsheets, the large number of data sets required for sound and convincing statistics necessitates a relational database system. Data in relational databases is usually stored in a multitude of tables, each containing a distinct portion of the data, together with information about how the entries of the diverse tables are related to each other. Database systems frequently used in science are the open source database systems MySQL (http://dev.mysql.com/) and PostgreSQL (http://www.postgresql.org/). Importing data into a database, or retrieving data from the database, is commonly performed using executable scripts, which in molecular biology are most frequently written in popular scripting languages like Python (https://www.python.org/) or Perl (https://www.perl.org/).

There are a number of good reasons for the utilisation of a relational database for genotype-phenotype correlation studies. The data produced in a variety of experiments, like those performed in "-omics" experiments, is mostly available in flat files. For fast access and reliable storage, the produced data has to be filed into a coherent database structure. The generation of data subsets for analysis is supported through the relational database schema by providing convenient ways to filter the stored information. Data sets from subsequent analyses can also be sorted into the database, thus allowing both original data and analysis results to be superimposed on the database structure. This in turn provides means to link the stored data with supporting information, in this way easing the interpretation of the produced results through an "expert system".

A reasonable and obvious choice for a basic database conception would be a structure that reflects the genomic organisation of features (genes, transcriptional units, operons, proteins, regulatory elements, etc.). The annotation files for finished bacterial genome projects from one of the public sequence repositories (DDBJ [8], EMBL [9], GenBank [10] or RefSeq [11]) can be used for its creation. Parsing the information into a suitable table structure provides not only the sequence information and annotation for the required genomic features but also a wealth of links to additional external data sources. A subsequent update of the database contents with sequence and annotation data for recently finished

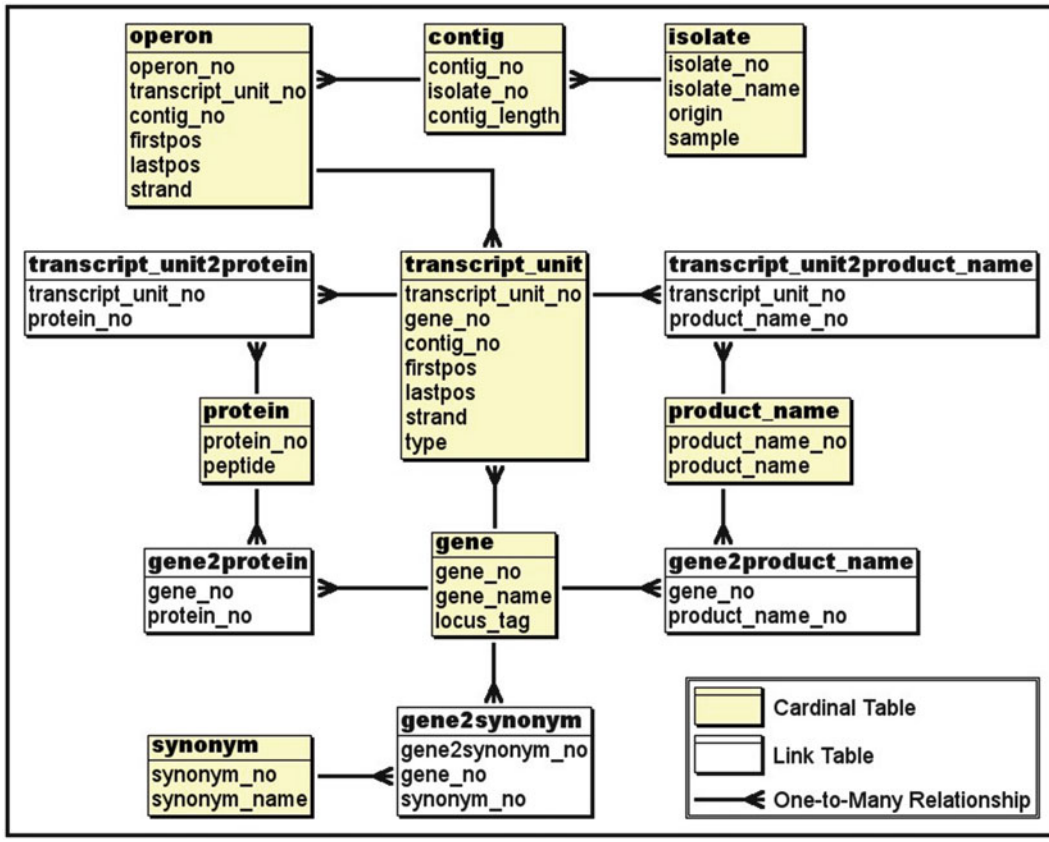


Fig. 1 Example of the basic structure for a relational genomic database. The tables reflect the genomic organisation. Additional tables for further structural genomic elements, phenotypic data and external database links have to be linked to the respective tables of the basic structure

strains of the bacterial species in question ensures that the data collection is up to date (Fig. 1).

The database system has to be extendable in a way that any additional experimental results of a multitude of different "-omics" technologies and supportive information can be easily integrated without requiring changes to the overall structure of the database. This also includes information defining the pan-genome (Sect. 2.3). With the genomic annotation as the basis of the data collection, and the "-omics" data, which represent molecular phenotypic properties of the respective prokaryote ("phenotypic landscape"), a broad data foundation for genotype-phenotype correlation studies is prepared.

Storing data in relational databases facilitates data filtering. For example in RNA-seq experiments data can be analysed based on isolate groups, read coverage or SNP scores/qualities. This facilitates the creation of data matrices for statistical methods, which can be filtered based on a multitude of parameters. The organisation of

data in a relational database system provides a suitable structure for the storage of the data produced by "-omics" experiments (absolute/relative expression values, overall variations in the expression of the trait, etc.). Furthermore, it enables the association of the data with additional information, like genomic feature annotation, or phenotypic data of interest. Phenotypic data might, for instance, include colony morphology parameters, minimal inhibitory concentration (MIC) profiles of antibiotic agents, growth curves, biophysical parameters of biofilm formation and so forth. With this method, a holistic information platform for data storage and data retrieval is available, which is not only a template structure for storage and integration of experimentally generated "-omics" data but also provides supporting information.

2.2 Data Visualisation and Distribution

A convenient way for users to communicate with a relational genomic database is a web-based system. Such a system can be easily implemented on a computer with a so-called LAMP configuration: **L**inux operating system, **A**pache web server, **M**ySQL database and **P**HP, Perl or Python scripting. A system with a browser-based user interface provides platform-independent user access not only locally but also globally; a password protection restricts access to authorised users. Web services creating dynamic web pages establish possibilities for data retrieval and data analysis and facilitate the interpretation of analysis results by connecting them with background information. The Common Gateway Interface (CGI) scripts required for the creation of the dynamically created web pages are frequently programmed in a popular scripting language like PHP (http://php.net/), Python or Perl.

2.3 Pan-Genome Creation

The starting point of any attempt to understand genotype-phenotype correlations is a detailed understanding of the genomic composition of the organism under consideration. To date, the full genomic sequence of multiple strains is available for many prokaryotes. Together with the annotation of genomic features this provides at least an initial understanding of the genetic organisation. The entirety of all different genes in an organism is called a pan-genome (or supra-genome) [12]. The concept describes the gene pool that is available for a given organism: the larger the set of different genes and gene functions which an organism has collected in its pan-genome, the higher the potential of the organism to settle in a multitude of environmental niches.

A compilation of the genes of which a species' pan-genome is composed is beneficial – or even required – in several different ways to understand phenotypical differences. Sometimes, already the presence of a gene or a group of genes in the accessory genome (the genes which are present in only a subset of strains) is sufficient to explain a given phenotype or its absence. A pan-genome can also be used in situations in which a reference genome is required for an

analysis, for instance, for base calling and transcriptional profiling (Sect. 2.4). Normally, the genome assumed to be the ancestral genome of the isolate under consideration is used as a reference. In some situations, however, the ancestral genome is unknown. Instead of arbitrarily choosing one of the known strains to be the reference, a more accurate approach is to use the pan-genome of the species, thus avoiding possible artefacts arising from the erroneous selection of the reference genome.

To identify the pan-genome for a species, the orthologous gene relationship between the genomic sequences of all strains, for which sequence and annotation are available, is determined. An orthologous relationship is defined as the reciprocal best hit of two genes not belonging to the same strain. An orthologous gene group is a group of genes in which all genes are orthologs of all other genes; if the gene group contains a gene from all investigated isolates, the respective gene group is a core gene group. Many ortholog gene groups, however, will contain only representatives of a subset of the investigated isolates (accessory genes). Some of the genes do not form a single ortholog gene pair; these are identified as singleton gene. For each of the orthologous gene groups a representative sequence is selected; all representatives are then compiled into a pan-genome gene list. Singleton genes are subsequently added. A number of online tools and software suites have been published, which assist in the creation of a pan-genome [13, 14].

The reference genome utilised for the evaluation of an RNA-seq experiment is used in the form of a FASTA file. In the case of a pan-genome the collected gene sequences are listed in the file, thus forming a template for read mapping and SNP detection. However, the sequence for a particular gene must not simply contain the annotated gene sequence (between start and stop position), because reads produced in a sequencing project frequently continue into the intergenic region. For this reason, gene sequences in the pan-genome should contain flanking sequence, so that even reads continuing into the intergenic region can be properly assigned. In the case of reference sequences created for DNA-resequencing experiments, the intergenic sequences must also be included in the construction of the pan-genome sequence. The intergenic nucleotides on both sides of a gene, but particularly upstream, are of interest, because they may contain regulatory elements of the gene in question.

2.4 RNA-Seq Experiments

In the first step of an RNA-seq [6] experiment, RNA is prepared from harvested cell material. Since less than 5% of prokaryotic cellular RNA is composed of mRNA sequences, rRNA and tRNA are depleted in order to improve the ratio of mRNA to rRNA and tRNA. Common techniques for mRNA enrichment in prokaryotes are ribosomal RNA capture, degradation of processed RNA and

selective polyadenylation of mRNAs [15]. From the RNA preparation, a cDNA library is created [16], which is subsequently sequenced.

For transcriptional profiling the reads produced in the RNA-seq experiment are matched to the pan-genome (Sect. 2.3) or a reference genome, resulting in a reads per gene (RPG) count. To do this, short reads of the sequence in question are aligned with the reference sequence. A multitude of read mapping software tools are available, two popular examples are Stampy [17] and Bowtie [18]. The calculated RPG can be used as the basis for the estimation of relative expression values by comparing them to the RPG of a reference strain; such normalisation is required for the comparison of the gene expression in different isolates. Those reads not matched during read mapping are not necessarily artefacts; a de novo assembly can be performed in an attempt to detect unique genes which were previously unknown (Sect. 2.6).

For the estimation of sequence variations ("SNP calling"), a reference genome is required. Normally a previously sequenced strain is used, which is ideally the ancestral genome of the variant under investigation. If the selection of the proper reference strain is arbitrary or inaccurate, alternatively a pan-genome can be used. In the SNP calling process, sequence differences are determined by matching the reads produced during sequencing of the isolate in question to the reference genome, for example using SAMtools [19]. The analysis of RNA-seq data with SNP calling algorithms helps to understand the overall variability on the single nucleotide level. This knowledge is useful in phenotype correlation studies, because it supplies required information for the weighting of a given SNP's contribution to the expression of the phenotype.

In comparison with DNA sequencing (DNA-seq), the sequence information gained through RNA-seq experiments is incomplete, because intergenic regions are in most parts not covered by the sequencing experiment. Nevertheless, through short read alignment of the high-quality sequencing read collection to the reference strain, information about sequence differences (SNPs, InDels) of the sequenced strain in comparison to the selected reference sequence is available. The big advantage of RNA-seq, however, lies in the fact that the experiment also provides information about relative gene expression, which is measured by the read coverage of a specific location of the genomic sequence. This provides information about genes that are active in the respective isolate, when it is cultivated under the applied growth conditions.

2.5 De Novo Genome Assembly

The reads resulting from DNA-seq can be mapped to a reference genome to assemble them into contigs. Although this is a faster approach to assemble contigs than a de novo assembly, the latter approach is unbiased. A de novo assembly of genomic reads can be done using tools like Velvet [20] or ABySS [21] with parameters

optimised to maximise the average contig length. Smaller contigs are removed, and larger contigs are assembled into supercontigs, for instance using the software CAP3 [22]. A reasonable threshold value for the distinction between small and large contigs is 200 bps. Supercontigs and unassembled regular contigs are annotated by matching them to all bacterial entries in the UniProt database [23] using the BLAST tool blastx [24, 25] to identify known proteins or genes from other strains or species. Results are filtered for the best hit for a given region in a contig in order to remove multiple descriptions for the same gene. BLAST matching of the resulting gene list against the pan-genome distinguishes known genes from yet unknown or undescribed genes.

2.6 De Novo Transcriptome Assembly

In the first step of de novo transcriptome assembly, the reads resulting from an RNA-seq experiment are mapped against a reference genome using tools like BWA [26], or Stampy [17], resulting in the identification of the associated genes. Those reads which do not map, however, still contain valid and important information and are not necessarily artefacts resulting from poor sequencing reads: if the sequenced isolate contains a gene which is not part of the reference genome, its sequencing reads will not match.

Catching these accessory genes is achieved by processing the pool of unmatched reads. In the first step, the mapping of the yet unmatched reads is performed against all strains of the species, for which sequence and annotation is available, or against the pan-genome if it has been previously created. This allows identification of all genes which are not part of the reference gene set but which are present in one of the species' other strains. With the remaining pool of sequencing reads a de novo transcriptome assembly approach using Velvet [27, 28] is performed. To start, an appropriate parameter set uses a minimal contig length of 100 bp and a sequence similarity higher than 90% as cut-off values, with application of a range of k-mer values (for example 27–37). As for de novo genome assembly, the resulting contigs are matched to the bacterial entries of the UniProt database [23] to annotate detected genes.

2.7 Phylogenetic Trees

Genotype-phenotype correlation studies rely upon statistical methods. For this reason, the number of isolates used for such a study is critical; a higher number of strains used in the study results in higher accuracy for the results of a statistical analysis. This statement, however, is not globally true – a higher number of isolates increases the significance of the applied statistical tests but only if they are properly selected.

The isolates collected for studies that want to explain genotype-phenotype correlations have to be collected from a broad spectrum of resources, to assure that they are genetically as diverse as possible. If, for instance, samples of pathogens were collected in a single hospital, there is a high risk that at least some of them are clonal

lineages caused by a clonal outbreak. These strains will likely carry almost identical genomic sequences; if isolates resulting from such an outbreak are included in a statistical analysis, although the overall number of genomes used in the study might seem appropriate, the results will be biased and could be misleading.

Prior to the start of a statistical project, the bacterial isolates used in the study should therefore be checked for independence. This can be done by the construction of a phylogenetic tree [29], in which clonal outbreaks will be easily detected. The construction of phylogenetic trees is based on a set of core gene sequences. For each of the isolates to be included in the phylogenetic tree, information for each of the selected genes is extracted. This can be achieved, for instance, by first creating a consensus sequence from the reads produced for the respective isolate using the SAMtools mpileup tool [19]. The locations of the marker genes are estimated by sequence matching with the reference gene sequences, whereby the isolate gene sequences are cut out and concatenated, typically using a short spacer nucleotide sequence consisting of "N"s to separate them. The result of the operation will be a collection of sequences, one for each of the isolates, each containing the concatenated sequences of the marker genes, always in the same order and orientation.

In many projects in which phylogenetic trees are constructed, a small number of housekeeping or marker genes are used. From the multi-sequence FASTA file created as described above, with a proper tool like ClustalW [30], a Phylip distance matrix and a phylogenetic tree can be constructed and subsequently visualised, e.g. in R statistics by using the as.dist, hclust and as.dendrogram tools.

The rather small collection of marker genes used for the construction of the phylogenetic tree, however, has the potential to result in a biased phylogenetic tree. Using a large number of genes for tree construction would be a much better choice, yet causes problems during the sequence alignment step due to huge computational costs and extremely long programme run times. An alternative is an alignment-free genome phylogeny method, for instance the k-mer tree method as described in Leekitcharoenphon et al. [31]. The basic idea behind the method is that highly similar sequences share k-mers (nucleotide sequence fragments of length k). The frequency of all k-mers across the genomes is computed and used to construct a matrix with k-mers as rows and genomes as columns, with the cells containing the k-mer frequency. This matrix is then submitted to hierarchical clustering in order to build the k-mer tree, e.g. as a neighbour-joining tree. With the k-mer tree method, all core genes which have a reasonable coverage (e.g. at least 90% of the gene) for all of the genomes can be used as input for the construction of the phylogenetic tree, thus using a huge number of genes for the construction of the tree. The length of the k-mers depends on the genome set and has to be adapted for best results.

2.8 Statistical Analysis of Variations Between Groups of Isolates

The investigation of complex behaviour, like resistance against antibiotic agents or the virulence of bacterial pathogens, requires the application of statistical methods. A strategy for the statistical analysis of "-omics" data sets for potential genotype-phenotype correlations is the comparison of two isolate groups, where one is positive with respect to a phenotypic trait and the other is negative with respect to that trait. The overall aim is to exploit the wealth of acquired data and to use the information in order to identify (groups of) sequence variations that reliably differentiate the groups, in order to understand the molecular mechanism characterising the phenotype and to identify biomarkers. For the statistical analysis of the collected data as well as the graphical representation of results, the open source statistical language R (http://www.r-project.org/) or the commercial available programme MatLab (http://www.mathworks.com/products/matlab/) can be used.

One way to analyse sequence variations is to consider each SNP or InDel individually, investigating their potential to be significant for the explanation of the group differences. For the evaluation of SNPs and InDels, a table is extracted from the relational database, which contains columns for the strains or isolates, a row for every SNP and the data matrix that is used as the input for statistical analysis. The cells of the table, the junctions of SNP rows and isolate columns, contain a 1 (for present), 0 (for absent) or NA (for unknown or insecure). Additional leading columns in the table contain information about the SNP itself, like the position of the respective SNP in the reference genome, intergenic or gene location of the SNPs and further information. This extra information is not required for the statistical analysis but rather to understand the results of the statistical test, for which Fisher's Exact Test (Fig. 2) is a commonly used algorithm.

Another approach assumes that any (non-synonymous) SNP or InDel in a given gene has an impact upon that gene, influencing its activity or even completely destroying it. As a consequence, the respective gene has significance for the explanation of the group differences. Again, a data matrix is extracted from the database, with columns for strains/isolates, and leading columns which contain additional information about the involved genes, but which are not required for the statistical test itself. The rows, however, this time stand for the genes: the cells of the table contain a 1 (for present) if the respective gene for the isolate in question has at least one SNP, a 0 if no SNP is present and NA if the sequencing results do not permit any definite assertion about the state of the gene. Again, Fisher's Exact Test is frequently the method of choice.

For expression values, Student's *t*-test (Fig. 3) is a possible statistical test to be applied. In the case of the example R script presented in Fig. 3, the input is a table file containing RPG counts for all isolates which are included in the analysis and an annotation

```
calculate.raw.p.values <- function(boolean.matrix, no.plus) {
  p.values <- c()
  for (i in 1:nrow(boolean.matrix)){
    pi <- 1
    if (sum(is.na(boolean.matrix[i,])) < ncol(boolean.matrix)) {
      no.true <- sum(boolean.matrix[i,], na.rm=T)
      no.false <- sum(!boolean.matrix[i,], na.rm=T)
      if (no.true > 0 & no.false > 0) {
        no.true.plus <- sum(boolean.matrix[i,1:no.plus], na.rm=T)
        no.false.plus <- sum(!boolean.matrix[i,1:no.plus], na.rm=T)
        ft <- fisher.test(matrix(c(no.true.plus, no.false.plus, no.true -
                          no.true.plus, no.false - no.false.plus), nrow=2, byrow=T))
        pi <- ft$p.value
      }
    }
    p.values <- c(p.values, pi)
  }
  return(p.values)
}

bh.correction <- function(p.values) {
  psort <- sort(p.values, index.return=T)
  p.values.bh <- p.values
  previous.p.value <- 0
  for (i in 1:length(p.values)) {
    p.values.bh[psort$ix[i]] <- p.values[psort$ix[i]]*(length(p.values)+1-i)
    if (p.values.bh[psort$ix[i]] > 1) {p.values.bh[psort$ix[i]] <- 1}
    if (p.values.bh[psort$ix[i]] < previous.p.value) {
                    p.values.bh[psort$ix[i]] <- previous.p.value}
    previous.p.value <- p.values.bh[psort$ix[i]]
  }
  return(p.values.bh)
}

matr <- read.table("./Input.mat", header=F, sep=";")
pv <- calculate.raw.p.values(matr, 6)
pv.bh <- bh.correction(pv)
write.table(pv.bh, file="./Output.txt", col.names=F, row.names=F, sep=";")
q("no")
```

Fig. 2 Example R script for Fisher's Exact Test (S. Pohl and F. Klawonn, HZI, Germany). The input file "Input.mat" contains semicolon-separated data, each line stands for a gene or a SNP, and each column represents an isolate. No header line is used. The fields of the matrix contain a 1 if a condition in question is fulfilled (e.g. if a given SNP is present in a given isolate), a 0 if the condition is NOT fulfilled and NA if the condition in question is ambiguous. The output file "Output.txt" contains a *p*-value for each line of the input file, in the same order

file from which the R script reads gene length information. A function contained in the DESeq package is used for the calculation of normalised gene expression values. The statistical analysis creates a list of genes, which are potentially involved in the cellular mechanisms that produce the observed phenotype.

It should be emphasised that although considerable information can be gained through statistical tests, the statistical analysis of differences between groups of isolates is limited in its potential to

```
require(DESeq)
rpgtab <- read.table("./rpg.tab", header=TRUE, row.names=1)
frstgrp <- c("B197", "B214", "B271", "B337", "B428", "B445")
scndgrp <- c("MHH10660", "MHH10978", "MHH11148", "MHH11540", "MHH11572", "MHH11785")
bothgrp <- c(frstgrp, scndgrp)
rpgdat <- rpgtab[, colnames(rpgtab) %in% bothgrp]
namelist <- colnames(rpgdat)
cdsdata <- newCountDataSet(rpgdat, namelist)
cdsdata <- estimateSizeFactors(cdsdata)
normrpg <- counts(cdsdata, normalized=TRUE)
annot <- read.table("./PA14_annot.tab", header=TRUE, row.names=1)
nrpkdat <- 1+normrpg/annot$length*1000
lnrpkdat <- log2(nrpkdat)
lnrpkdat <- lnrpkdat[, bothgrp]
t.test <- apply(lnrpkdat, 1, function(lnrpkdat) {
       t.test(x=lnrpkdat[1:6], y=lnrpkdat[7:12])$p.value})
p.adjust <- p.adjust(t.test, method="fdr")
t.test_results <- cbind(lnrpkdat, t.test, p.adjust)
frst_lnrpk <- lnrpkdat[, colnames(lnrpkdat) %in% frstgrp]
scnd_lnrpk <- lnrpkdat[, colnames(lnrpkdat) %in% scndgrp]
frst_median <- apply(frst_lnrpk, 1, median)
frst_sd <- apply(frst_lnrpk, 1, sd)
scnd_median <- apply(scnd_lnrpk, 1, median)
scnd_sd <- apply(scnd_lnrpk, 1, sd)
non_medians_sds <- cbind(frst_median, frst_sd, scnd_median, scnd_sd)
meddiff <- non_medians_sds[, 1]-non_medians_sds[, 3]
non_medians_sds <- cbind(non_medians_sds, meddiff)
colnames(non_medians_sds) <- c("Frst Med", "Frst SD", "Scnd Med", "Scnd SD", "Med Diff")
write.table(non_medians_sds, "./Output.non-med", sep="\t")
write.table(t.test_results, "./Output.t-test", sep="\t")
q("no")
```

Fig. 3 Example R script for Student's *t*-test (S. Pohl and F. Klawonn, HZI, Germany). The data file "rpg.tab" (Reads per Gene) contains lines with<TAB>-separated fields, in which a leading column provides the gene names and each following column represents an isolate. The first line is a header line and contains the isolate names. Data file "PA14_annot.tab" contains<TAB>-separated PA14 gene data. Each line describes a gene; the order of the genes has to be the same as in "rpg.tab". The first line of the file is a header line describing the contents of the respective column. For the above script, only one column is important: it is called "length" and contains the length of the respective gene in nucleotides. The gene length is required for read count normalisation

provide satisfactory results. For all tests described above, the results are lists of sequence variations, sorted according to the calculated *p*-values. They form the basis for the development of a working hypothesis for the explanation of the molecular mechanisms that lead to the observed phenotypic trait. This working hypothesis then is the starting point for wet lab experiments performed to prove or reject the hypothesis. An issue that always has to be considered is the *p*-value cut-off that has to be applied to yield a minimal rate of false positives, where sequence variations are detected as significant although they are not, and false negatives, where sequence variations that are actually significant are not present in the results list of significant features. An appropriate means for the definition of a reasonable cut-off value utilises permutation tests. By randomly

shuffling the contents of the cells within the rows of the data matrix relevant for statistical analysis, a randomised data set is created. The statistical analysis of this arbitrary data set provides random *p*-values; the lowest *p*-value detected in a set of permutation tests is a reasonable cut-off value to exclude insignificant results.

References

1. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402
2. Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem 6:287–303
3. Bielecki P et al (2014) In vivo mRNA profiling of uropathogenic *Escherichia coli* from diverse phylogroups reveals common and group-specific gene expression profiles. mBio. doi:10.1128/mBio.01075-14
4. Pohl S et al (2014) The extensive set of accessory *Pseudomonas aeruginosa* genomic components. FEMS Microbiol Lett 356:235–241. doi:10.1111/1574-6968.12445
5. European Commission (2010) Workshop to clarify the scope for stratification biomarkers and to identify bottlenecks in the discovery and the use of such biomarkers. http://ec.europa.eu/research/health/pdf/biomarkers-for-patient-stratification_en.pdf. Accessed 19 Mar 2015
6. Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63
7. Schulz S et al (2015) Elucidation of sigma factor-associated networks in Pseudomonas aeruginosa reveals a modular architecture with limited and function-specific crosstalk. PLoS Pathog 11:e1004744. doi:10.1371/journal.ppat.1004744
8. Tateno Y et al (2002) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 30:27–30
9. Kulikova T et al (2004) The EMBL nucleotide sequence database. Nucleic Acids Res 32: D27–D30
10. Benson DA et al (2014) GenBank. Nucleic Acids Res 42:D32–D37
11. Pruitt KD et al (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42:D756–D763
12. Medini D et al (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594. doi:10.1016/j.gde.2005.09.006
13. Vernikos G et al (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154. doi:10.1016/j.mib.2014.11.016
14. Xiao J et al (2015) A brief review of software tools for pangenomics. Genomics Proteomics Bioinformatics 13:73–76. doi:10.1016/j.gbp.2015.01.007
15. Sorek R, Cossart P (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat Rev Genet 11:9–16
16. Head SR et al (2014) Library construction for next-generation sequencing: overviews and challenges. Biotechniques 56:61–77
17. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936–939. doi:10.1101/gr.111120.110
18. Langmead B et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi:10.1186/gb-2009-10-3-r25
19. Li H et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
20. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. doi:10.1101/gr.074492.107
21. Birol I et al (2009) De novo transcriptome assembly with ABySS. Bioinformatics 25:2872–2877. doi:10.1093/bioinformatics/btp367
22. Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877
23. UniProt Consortium (2014) Activities at the universal protein resource (UniProt). Nucleic Acids Res 42:D191–D198. doi:10.1093/nar/gkt1140
24. Altschul SF et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410
25. Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
26. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler

transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324

27. Zerbino DR (2010) Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics. doi:10.1002/0471250953.bi1105s31
28. Zerbino DR et al (2009) Pebble and Rock Band: heuristic resolution of repeats and scaffolding in the Velvet short-read assembler. PLoS One 4:e8407. doi:10.1371/journal.pone.0008407
29. De Bruyn A et al (2014) Phylogenetic reconstruction methods: an overview. Methods Mol Biol 1115:257–277. doi:10.1007/978-1-62703-767-9_13
30. Larkin MA (2007) ClustalW and ClustalX version 2. Bioinformatics 23:2947–2948. doi:10.1093/bioinformatics/btm404
31. Leekitcharoenphon P et al (2014) Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica. PLoS One 9:e87991. doi:10.1371/journal.pone.0087991

Systems Biology Tools for Methylotrophs

Marina G. Kalyuzhnaya, Song Yang, David A.C. Beck, and Ludmila Chistoserdova

Abstract

The methylotrophy field is currently experiencing a renaissance. Innovative cultivation techniques are resulting in discovery of novel types of methylotrophs, the growing genomic databases are providing blueprints for metabolic reconstruction in traditional as well as newly discovered methylotrophs, and the concepts and dogmas formed during the pre-omics era are changing, sometimes dramatically. The emerging approach in characterizing methylotrophs, as well as other metabolic specialists, is a combination of systems biology approaches, with availability of the genomic sequence being a prerequisite. We here describe a series of omics approaches to characterizing methylotrophs, which, in their combination, provide comprehensive outlook at how methylotrophy metabolism is enabled in specific methylotroph guilds and how it is regulated.

Keywords: Community dynamics, Genomics, Metabolic modeling, Metabolomics, Methanotroph, Methylotroph, Proteomics, Transcriptomics

1 Introduction

Methylotrophs are microbes that can build their biomass as well as obtain energy from compounds with no carbon–carbon bonds (C1 compounds), which include methane, methanol, methylated amines, methylated sulfur species, and halogenated C1 compounds [1]. Aerobic methylotrophs have been known since early 1900s [2], with the details of the biochemical pathways responsible for utilization of C1 compounds emerging in the early 1970s [3]. Anaerobic methylotrophy has been described only recently. While in Bacteria it appears to be linked to nitrate/nitrite reduction [4], in Archaea, it can be linked to nitrate reduction [5], sulfate reduction [6], or reduction of certain metals [7]. None of the anaerobic methylotrophs have yet been cultivated in pure cultures. The focus of this chapter is on methylotrophic bacteria and does not cover the archaeal methylotrophs.

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 97–118, DOI 10.1007/8623_2015_69, © Springer-Verlag Berlin Heidelberg 2015, Published online: 10 April 2015

2 Genomics

Today, a (draft) genomic sequence is prerequisite for studying aspects of physiology and metabolism of a select model organism. Typically, genomic sequencing, including gene calling and gene annotation, is outsourced to service facilities, such as the Joint Genome Institute (http://www.jgi.doe.gov), Genoscope (http://www.genoscope.cns.fr), etc. These facilities have generated most of the genomic sequences for methylotrophs to this date, and they deliver high-quality genome drafts or finished genomic sequences [8, 9]. Gene annotations can be manually curated by experts, based on experimental evidence [10]. Of course, genomic sequencing, assembly, and annotation could be carried out in the house, using one of the current sequencing platforms and appropriate software packages [11, 12].

3 Functional Metagenomics

3.1 Principle

While genomics of methylotrophs are no different from genomics of any other type of microbes, special metagenomic approaches are effective in studying methylotrophs in semi-in situ conditions, such as "functional" metagenomics combining stable isotope probing (SIP) with high-throughput sequencing. The technique of SIP, as applied to probing active methylotrophs in situ, was first developed and further perfected in the Colin Murrell's laboratory, and we refer the reader to their excellent reviews as well as to the original works (e.g., [13–16]). This approach involves feeding natural populations a substrate of interest, labeled by a heavy isotope (e.g., ^{13}C), followed by characterization of the heavy fraction of communal DNA that should be enriched in DNA of microbes that actively metabolize the labeled substrate. A number of labeled C1 substrates are commercially available.

3.2 Sample Collection and Cell Labeling

Samples are collected using an appropriate devise, such as box core, Niskin bottles, etc., and transported to the laboratory on ice. Microcosms are set up in conditions that mimic the in situ conditions, including substrate concentrations that should approximate the in situ concentrations. However, substrate concentrations should be high enough to allow for efficient labeling. We were successful using the following concentrations of ^{13}C-labeled substrates: methane (up to 50% of the atmosphere) [17, 18], methanol (1–10 mM) [17, 19], methylamine (10 mM), formaldehyde (1 mM), and formate (10 mM) [17]. Samples are incubated, preferably at the in situ temperature, for the duration of time that allows for some of the DNA to become labeled with ^{13}C, 3–30 days, depending on the substrate and on other conditions [17–19].

3.3 DNA Extraction, Isopycnic Centrifugation, and Labeled DNA Recovery

Total DNA is extracted using an appropriate protocol for a given sample. PowerSoil DNA Isolation Kit (MO Bio Laboratories, Carlsbad, CA, USA) produces good results with soil and sediment samples. DNA is prepared for CsCl-ethidium bromide density gradient ultracentrifugation as previously described [13] and centrifuged at approximately 265,000 g (e.g., in Beckman VTi 65 rotor) for 16 h at 20°C. DNA fractions are visualized in UV, and ^{13}C-DNA fractions are collected using 19-gauge needles [20]. DNA preparations may be subjected to a second round of CsCl-ethidium bromide density gradient ultracentrifugation, followed by a standard DNA purification procedure. Optionally, the ^{12}C-DNA fractions may also be collected and analyzed, for comparison.

3.4 DNA Sequencing and Assembly

Shotgun libraries are constructed in accordance with the protocols specific to each sequencing technology, following manufacturers' suggestions. Illumina is probably the most attractive technology currently, providing the optimal cost efficiency per sequenced nucleotide [21]. Like sequencing single genomes, metagenomic sequencing is also best outsourced to specialized facilities. However, respective pipelines can be developed within a single laboratory. Raw sequences are trimmed and denoised, paired-end sequences are joined, sequences are assembled, and genes are called and functionally annotated using appropriate software for each step. Service facilities such as the JGI have developed efficient pipelines for processing the data, and these are constantly updated and improved, and we refer the reader to their expertise (http://img.jgi.doe.gov/cgi-bin/m/main.cgi) [22]. Sequence coverage and degree of assembly depend on the sequencing effort applied and on the species richness and evenness of the enriched communities.

3.5 iTag (Pyrotag) Sequencing

These days it is customary to carry out phylogenetic profiling of community DNA via iTag (Illumina) or Pyrotag (Roche 454) sequencing, targeting variable regions of the 16S rRNA gene, after polymerase chain reaction (PCR) amplification [23, 24]. Such profiling provides an estimate of the complexity of the community to be analyzed via metagenomics, suggesting, importantly, an appropriate sequencing effort for the metagenome in question. The sequencing effort for iTagging (Pyrotagging) may vary between tens and hundreds of thousands sequencing reads per sample, allowing for high phylogenetic resolution, even for very complex community samples. These days, the experimental part of this approach (PCR amplification, sample purification, and sequencing), as well as statistical analyses, could also be outsourced to service facilities. For fast turnaround, we use MR DNA service facility (http://www.mrdnalab.com/). i/Pyrotagging is also efficient in monitoring enrichment for specific functional types in the functional metagenomics experiments. For example, Pyrotag

profiling of Lake Washington sediment communities enriched in heavy carbon isotope originating from methane has revealed not only rapid reduction in community complexity, but also significant enrichment in the sequences of *Methylococcaceae* and *Methylophilaceae*, suggesting, on one hand, that they are key players in methane consumption and, on the one hand, that they might be involved in cooperative behavior [18].

3.6 Data Analysis

Several software packages are available for 16S rRNA gene amplicon sequencing, including QIIME [25], MOTHUR [26], and MEGAN [27]. In addition, well-documented protocol and software combinations are available for simplifying typical analyses of microbial communities based on 16S rRNA gene sequences [28]. At their core, these tools perform a very similar set of sequence handling, clustering, and taxonomy assignment steps.

3.6.1 Sequence Handling

When paired-end Illumina reads are used, the two read pairs need to be assembled to create the full-length sequence. In this step, there are critical parameters that need to be evaluated for each experiment, which include how many bases of an overlap are required between the read pairs and how many mismatches are allowed in the alignment. For both Illumina- and Roche 454-based reads, the full-length sequences can be trimmed. In our experience, the maximum error method of Edgar has proven useful [29]. In this approach, each read is scanned from 5′ to 3′ using the sequencer-generated quality scores, to estimate the cumulative number of sequencing errors. A maximum expected error of 0.5 has the effect of limiting the number of sequence errors to less than 1. Larger numbers result in longer sequences, whereas smaller numbers result in shorter but more accurate sequences.

3.6.2 Sequence Clustering

The quality trimmed reads are then run through preprocessing steps. A good summary can be found in [29]. Briefly, unique sequences are counted and checked for chimeras that can result from amplification. Chimera detection can be done in a variety of ways including the UCHIME/USEARCH approach against a high-quality reference database [30] or the ChimeraSlayer tool [31].

Next, the sequences are clustered at a given percent identity, from 95% to 99%. Tradeoffs between artificial splitting of clusters due to possible amplification errors at high identities and species resolution at the lower identities are a necessary aspect of the clustering process. For each cluster, a representative sequence must be chosen or computed (e.g., average) for subsequence taxonomical assignment. Representative sequences have the advantage of being real sequences but may not accurately represent an entire cluster at lower percent identity cutoffs, whereas average sequences may not be realistic and can frustrate assignment.

3.6.3 Taxonomy Assignment

Representative of each cluster is assigned to a taxonomical clade, which is attributed to the entire cluster. Assignment can be done using sequence alignment or another classifier method such as the Ribosomal Database Project (RDP) Classifier (rdp.cme.msu.edu) [32]. The RDP Classifier has the advantage of providing confidence scores in the assignment at each level of taxonomy.

3.6.4 Visualization and Analysis

Once the taxonomies have been assigned, a variety of visualizations and analyses can be performed after normalization. Normalization for sample size is required to assign relative abundances across samples when the total number of reads passing quality control varies between samples. Visualizations such as heatmaps and column charts of taxonomy are useful, particularly when only the most abundant OTUs are presented. That is, only those OTUs present in at least one sample at, for example, 1% or 2.5% are included. Tools such as *vegan* can be used to create ordination plots and perform principal component analysis and classical correspondence analysis [33].

3.6.5 Phylogenetic Markers in Metagenomes

To classify the 16S rRNA gene sequences in the metagenomes, 16S rRNA genes identified as part of an annotation pipeline are aligned against the RDP Classifier, as above. The top scoring alignment for each sequence is used to assign taxonomy, based on the annotations in the RDP.

3.6.6 Functional Gene Profiling

Enrichment for specific functional genes can be addressed in a similar way, via single-gene profiling, using known genes/proteins as queries in BLAST analyses [17, 18]. This approach is especially effective in the case of shallow metagenomes. We employed proteins involved in the reactions of the tetrahydromethanopterin pathway for C1 transfers that is a hallmark pathway in methylotrophs [34] to demonstrate the function-relevant enrichment of microcosm datasets [17, 18]. In a similar fashion, specific functions in methylotrophy can be profiled, such as soluble versus particulate methane monooxygenase [35, 36], MxaFI- versus XoxF-type methanol dehydrogenase [35–37], as well as the nitrogen metabolism functions, such as nitrogen fixation and assimilatory versus dissimilatory denitrification [18]. In cases of highly divergent enzymes, multiple queries are required. For example, we used peptide sequences of *fae* homologs belonging to different phylogenetic groups (Proteobacteria, Planctomycetes, and Archaea) to identify multiple and extremely divergent *fae* and *fae*-like sequences in our datasets [17, 18].

3.6.7 Assembling Genomes from Metagenomes

Genomes of individual organisms or populations of closely related strains may be present in a metagenome at high sequence coverage. Estimates of coverage for each organism can be initially gained from the coverage of individual 16S rRNA genes present in a

metagenome, as described above. If high relative abundance is predicted for a specific organism, it is reasonable to assume that complete or nearly complete genomes of respective strains may be present in a metagenomic dataset, and these can be extracted using one of the available binning tools. In the metagenomic study of Lake Washington methylotroph populations, a composite genome of *M. mobilis* totaling slightly over 11 Mb was extracted from the methylamine microcosm metagenome using a compositional binning method PhyloPythia, and genome completeness was validated by examination of the presence of key metabolic and housekeeping genes [17]. With satisfactory results, metabolic potential of an individual organism or a population of closely related strains (which will not necessarily be distinguished between by the binning methods) can be reconstructed, and genome-wide comparisons may be carried out with other complete or composite genomes. For example, by comparing the composite genome of *Methylotenera mobilis* to the complete genome of a close relative *Methylobacillus flagellatus*, we were able to uncover examples of highly conserved metabolic traits, including methylotrophy, as well as of non-conserved metabolic traits, including nonhomologous replacements in common biochemical pathways, such as a cytochrome electron acceptor from methylamine dehydrogenase in *M. mobilis* versus an azurin protein in *M. flagellatus* [17].

3.6.8 Genome Recruitment and Variant Analysis

The content of the metagenomes can also be evaluated via genome recruitment, as long as a reasonably closely related proxy genome is available for comparison. This approach is useful for less well-covered genomes, shorter reads, or single-end sequence-based experiments where binning and/or assembly may be problematic. In this mode, paired-end or single-end raw reads are aligned to a proxy genome at a given identity cutoff, e.g., 90–95%. Such relaxed alignment is available in BWA [38] when seeding is disabled (option –l 100,000). Similar modes are available for BFAST [39] and Bowtie 2 [40].

After the alignments are computed, SAMtools can be used to post-process the alignments to binary alignment file formats (.BAM) and to perform SNP calling [41]. A wide array of constantly evolving tools can also use the binary alignment file format (.BAM) for SNP predictions. Visualization of the alignments for manual inspection can be performed with IGV [42]. This protocol allowed us to assign up to 60% of reads in our relatively shallow metagenomes to specific organisms, not only at the genus and species level, but also at the ecotype level [43].

4 Community Dynamics Assessed via i/Pyrotagging

4.1 Principle

The active populations of methylotrophs in specific environmental niches can also be assessed without labeling, thus avoiding any biases that may be associated with this step, by following the dynamics of the communities in response to a specific stimulus. By profiling communities in the microcosms responding to the stimulus of methane, we were able to confirm our findings from SIP-metagenomics on the dominant role of *Methylobacter* in methane metabolism in the sediment of Lake Washington [43, 44]. In these experiments, we also obtained further support for the involvement of *Methylophilaceae* in metabolism of methane, by observing a dramatic increase in the population of *Methylophilaceae* in microcosms fed methane as the only carbon source [43, 44]. By modifying oxygen levels in these experiments, beside the *Methylotenera* species, we were also able to detect the response by the *Methylophilus* species, which were not detected via SIP-metagenomics [17, 18, 43, 44]. From the microcosm experiments, it appears that the *Methylophilus* species tend to be outcompeted by the *Methylotenera* species in hypoxic conditions, likely suggesting that microcosm incubations with ^{13}C-labeled substrates must have been limited by oxygen even when designated as "aerobic" [18]. This conclusion is supported by our recent data on active oxygen consumption by complex sediment communities [44].

4.2 Experimental Setup

Natural samples are collected using appropriate tools, such as box core, Niskin bottles etc., and transferred to the laboratory at appropriate temperature. Samples can be immediately used for setting up microcosm cultures or deep-frozen with a cryoprotective agent such as dimethyl sulfoxide (DMSO), to assure the viability of the microbial population and to preserve cells for omics profiling [45].

Microcosms are set up in a chosen set of conditions. Either water from the native environment can be used (e.g., above-the-sediment lake water in our case) or one of a traditional artificial media for methylotroph cultivation [46, 47], or a diluted version of the standard medium [43, 46]. Vials with rubber stoppers are used as cultivation vessels for methanotroph enrichments. Choosing the atmosphere composition in the headspace is important and should depend on the specific scientific questions. In natural environments such as lake sediments or wetlands, methane and oxygen are present as steep counter gradients [48, 49]. The specific methylotroph communities may inhabit specific microenvironments with different oxygen to carbon ratios [50]. In our experiments, we observed clear dependence of community compositions on the oxygen concentration [43, 44].

Microcosms are preferably incubated at an in situ temperature if insights into the natural processes are desired. Shaking is advised for

reproducibility. It is advisable to measure key chemicals, such as methane and oxygen concentrations, to monitor their consumption. Microcosm cultures can be sampled at the desired frequency by pelleting fractions of the culture, followed by DNA extraction, as above. DNA samples are then subjected to either phylogenetic profiling or to metagenomic sequencing using the methodologies described above.

5 Transcriptomics

5.1 Principle

While genomes and metagenomes provide the blueprint for an organism/community metabolic potential, the availability of transcriptomes, an outcome of the transcriptomics approach, brings them alive. Which genes are expressed in which conditions and how much more expression is needed to enable a certain lifestyle? Transcriptomics are also a major tool for discovering genes and functions that are involved in a specific metabolic scheme, but are not detectable by other types of experiments, such as mutant screening. Transcriptomics are most informative when transcript profiles between multiple growth conditions are compared. Dependent on the goal of the experiment and on the scientific questions pursued, samples can be collected from cultures growing on different substrates or in different conditions. While it is straightforward when it comes to facultative methylotrophs, it may be tricky with obligate methylotrophs, if only a single growth substrate is known to support growth. In the latter case, however, other conditions could be modified, for example, oxygen pressure or nitrogen source.

5.2 Sampling and RNA Isolation

The culture should either be grown in bioreactor (chemostat, turbidostat, or feed-batch) or grown in vials/flasks to the growth stage of interest (usually exponential phase of growth). At least two biological replicates should be used for each condition tested. Growth is terminated by the addition of 10% (V/V) of "stop solution." The "stop solution" is comprised of 5% water-equilibrated phenol (pH 6.6–7.0) and 95% ethanol. Cells are collected by centrifugation, typically at approximately 4,500 g at 4°C for 10–15 min. The resultant pellets are used for RNA extraction using RNeasy Kit (QIAGEN), essentially as described [51]. Higher RNA yields can be achieved by using a two-step procedure, as described [52, 53]. In this protocol, cell pellets are resuspended in 0.75 mL of the extraction buffer (2.5% CTAB; Sigma, St. Louis, MO), 0.7 M NaCl, and 0.075 M phosphate buffer (pH 7.6) and transferred into a 2-mL sterilized screw-cap tube containing 0.75 mL of phenol/chloroform/isoamylic alcohol with a volume ratio of 25:24:1 (Ambion, Austin, TX), 0.5 g of 0.1 mm silica beads (BioSpec products, Bartlesville, OK), 0.2% SDS (Ambion; Austin, TX), and 0.2% lauryl sarcosine (Sigma, St. Louis, MO). The

mixture is homogenized in a bead beater (Mini-Beadbeater, BioSpec Products; Bartlesville, OK) for 2 min (75% of the maximum power). The resulting slurry is centrifuged for 5 min at 4°C at approximately 20,000 g. The aqueous layer is transferred to a fresh tube containing 0.75 mL of chloroform/isoamylic alcohol with a volumetric ratio of 24:1 (Sigma, St. Louis, MO) and centrifuged again for 5 min at 4°C and 20,000 g, to remove the dissolved phenol. The aqueous phase is transferred to a new tube. $MgCl_2$ (final concentration 3 mM), sodium acetate (10 mM, pH 5.5), and 0.8 mL ice-cold isopropanol are added. Nucleic acids are precipitated at −80°C overnight. Samples are then centrifuged for 45 min at 4°C at 20,000 g, washed with 0.5 mL 75% ethanol (Deacon Labs, Inc.; King of Prussia, PA), and purified using RNeasy Kit (QIAGEN), essentially as recommended by the manufacturer.

The MICROBExpress™ Kit (Ambion, Austin, TX) or Ribo-Zero (Epicentre, Madison, WI) kits could be used to reduce the rRNA concentration, in order to increase sequencing depth for mRNA. Recently, such services are also best to outsource to service facilities, as part of a sequencing package. Alternatively, total RNA could be sequenced, with rRNA sequences removed from the dataset computationally.

Sample quality is monitored using three different techniques: (1) electrophoresis in TAE buffer in 1% agarose gel; (2) using Agilent 2100 Bioanalyzer with Agilent RNA 6000 Nano Kit, following suggestions by the manufacturer; and (3) real-time reverse-transcriptase PCR (RT-RT PCR) with appropriate 16S rRNA and/or functional gene-specific primers.

5.3 Transcript Sequencing, Alignment, and Read Mapping

The enriched or non-enriched RNA samples (at least two biological replicates) are typically submitted to a service facility equipped with up-to-date sequencing equipment such as Illumina, unless a sequencing machine is available at the premises and the expertise is in hand. As the technologies are constantly changing, the choice of the sequencing platform is normally dictated by cost per nucleotide.

5.4 Data Analysis

The resulting reads are aligned to the reference genome using an alignment tools such as BWA [38] BFAST [39], or Bowtie 2 [40]. The alignments are then post-processed into sorted BAM files with SAMTools [41]. Reads are attributed to open reading frames (ORFs) using the *htseq-count* tool from the "HTSeq" framework in the "intersection-nonempty" mod [54]. Samples can then be normalized by reads per kilobase of gene per million mapped reads (RPKM) to coding sequences [55] and resulting *p*-values from statistical comparisons corrected by a multiple hypothesis testing correction scheme such as *q*-values [56]. Alternatively,

normalization, comparison, and correction can be performed in a single package with the DESeq2 [57, 58].

6 Metatranscriptomics

6.1 Principle

Metatranscriptomics, as a natural next step to metagenomics, is what transcriptomics is to genomics. Thus, while metatranscriptomes may be analyzed in isolation, including assembly/binning and annotation, availability of matching (meta)genomic scaffolds make metatranscriptomics more reliable, most effective, and easier to process. These days it seems reasonable to combine analysis of metagenomes and metatranscriptomes from the same sample. Other than dealing with DNA and RNA originating from community samples, the strategies for sequencing and analysis are the same as described above. Again, depleting or not depleting total RNA of ribosomal RNA is a choice, dependent on what sequencing effort can be attempted. If total RNA is sequenced, then the rRNA sequences are simply computationally removed from the database before mRNA sequence mapping.

7 Proteomics

7.1 Principle

Proteomics, analysis of protein profiles of microbial cultures grown on specific substrates or in specific conditions, presents further opportunity to address the function directly, as proteins are the molecules that ultimately perform the function. Like transcriptomics, proteomics are most informative when transcript profiles between multiple growth conditions are compared, thus again becoming somewhat limiting when a single or few substrates support growth. Proteomics have been successfully applied in methylotrophs to define global proteome landscapes during methylotrophic versus non-methylotrophic growth [59–62], to obtain new insights into methylotrophy [63–65], to evaluate copper response [66], and to identify proteins produced specifically in the phyllosphere [59]. Different types of proteomics approaches have been applied, including analysis of select proteins, after two-dimensional gel electrophoresis separation [59, 62], or global (shotgun) proteomics [60, 61, 63–65]. The advantage of the latter technology is in its comprehensive nature. With sufficient number of spectra collected for a given proteome, up to 70% of the inferred proteome could be covered, and data from different conditions can be compared in a semiquantitative manner [60, 63–65]. Of course, to achieve this level of protein detection, a high-quality genomic sequence for an organism in question is required. Below we describe a typical workflow for shotgun proteomics.

7.2 Sampling and Protein Isolation

Cultures can be grown in a bioreactor or in vials/flasks, to the growth stage of interest (usually exponential phase of growth). At least two biological replicates should be used for each condition tested. Cells are cooled rapidly and harvested by centrifugation at approximately 4,500 × g for 5–10 min at 4°C. Cells are washed with cooled 15–20 mM Tris–HCl pH 8.0 buffer and flash-frozen in liquid nitrogen. Cells can be stored at −80°C for up to a few months.

Frozen cell pellets are resuspended in 500 μL hot resuspension buffer (20 mM Tris–HCl pH 8.0, 5 mM dithiothreitol (DTT)) and lysed by boiling for 2 min in a water bath, followed by cooling on ice for 10 min. For nucleic acid digestion, 10 units Benzonase nuclease (Roche) is added after adjusting the suspension to 2 mM $MgCl_2$ (final concentration). After 15-min incubation at room temperature, 200 μL of ethanol- and buffer-washed glass beads (150 μm) are added, and the sample volume is adjusted to 1.5 mL with resuspension buffer. Bead beating is performed for 4 min at 48 RPM, in a Mini-Beadbeater (BioSpec Products). Total protein concentration is determined by Bradford protein assay (Bio-Rad). The desired concentration is approximately 1 mg mL^{-1}. The homogenate is lyophilized to dryness. After resuspension in 500 μL digestion buffer (2 M urea, 5% acetonitrile, 5 mM DTT, 0.1% RapiGest), the sample is digested as follows. After reduction with 10 mM DTT for 30 min at 37°C and alkylation with 30 mM iodoacetamide for 30 min at 30°C (in the dark), digestion is started by adding EndoLysC (Roche Applied Sciences, Indianapolis, IN, USA) at a ratio of 1:200 protease to protein, according to the Bradford Assay, and the mixture is incubated overnight at 37°C. Trypsin (Promega, Madison, WI, USA) is then added at a ratio of 1:50, and incubation is continued for 8 h at 37°C. The reaction is stopped by adding 0.3% (final concentration) trifluoroacetic acid (TFA) until the pH reaches 2.5. RapiGest™ is precipitated by further incubation at 37°C for 30 min. The sample is then centrifuged for 10 min at approximately 20,000 × g, and supernatant is collected. Pellets (containing insoluble particles and the glass beads) are washed twice with 250 μL 0.5% TFA and 5% acetonitrile. pH is adjusted to 2.5 with TFA if necessary.

7.3 HPLC Pre-fractionation and Linear Ion Trap Mass Spectrometry

The soluble fraction after digestion is lyophilized to a volume of approximately 150 μL and centrifuged for 10 min at 20,000 × g. A 10–20 μL sample of the supernatant is applied to a PLRP-S reversed-phase column (2.1 mm i.d. × 150 mm, 300 Å, 5 μm; Polymer Laboratories) with mobile phases of 0.1% TFA in water and acetonitrile. Peptides are eluted at 0.2 mL min^{-1} with a gradient of 2–60% acetonitrile in 60 min and 60–90% acetonitrile in 20 min and collected as five separate fractions. The fractions are each lyophilized to 20 μL, and after reconstitution to volumes of 120 μL with acetic acid and acetonitrile at final concentrations of

0.5% and 5% (V/V), respectively, subjected to LC/LC-MS/MS using a biphasic 2-D capillary HPLC system (Michrom Magic 2002, Michrom, Auburn, CA, USA), coupled to an LTQ linear ion trap mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA). The peptides are eluted by ammonium acetate solutions (0, 10, 25, 50, 100, 250, and 500 mM), followed by reverse-phase gradients of 5–12% in 1 min, 12% B in 9 min, 12–40% B in 50 min, 40–80% B in 11 min, 80% B in 10 min, and 80–5% B in 5 min. Solvent B: 99.5% acetonitrile, 0.5% acetic acid (v/v). The MS^1 scan range is 400–2,000 m/z. After each main beam (MS^1) scan, the 10 most intense ions above threshold are selected for CID scans with one CID scan collected for each of the precursor ions. Default parameters under the Xcalibur 1.4 data acquisition software (Thermo Fisher) are used, with the exception of an isolation width of 3.0 m/z units and normalized collision energy of 40%.

7.4 Data Processing and Normalization

Proteomics data processing can be performed using entirely open source tools by leveraging the Trans-Proteomic Pipeline, TPP [67]. The spectra files can be converted to compressed mzXML files using the ProteoWizard's *msconvert* tool [68]. These spectra, in conjunction with a FASTA file containing the predicted amino acid sequences from the genome and several additional enzymes used in sample prep, including trypsin precursor (gi 136429), DNase I precursor (gi 6647483), RNase A precursor (gi 133198), and lysyl endopeptidase (gi 7463016), are fed into a peptide-spectra matching software such as COMET [69]. It is a good practice to use a set of decoy proteins that have the same size and amino acid composition as the proteome being searched. These decoys can be used to determine the false-positive rates arising from the peptide-spectra matching (PSM) process.

Next, the resulting pepXML from technical replicates are pooled and analyzed with PeptideProphet [70] using TPP's *xinteract*. During this step, the prefix used to identify the decoy proteins (e.g., prefix "DECOY" = *-OP-dDECOY*) is supplied with the pepXML and the FASTA. Additional options to consider include the confidence below which to ignore PSMs (e.g., 0.85 = *-p 0.85*) and the minimum length of a peptide to consider (e.g., 7 = *-l 7*). The resulting pepXML is then processed by ProteinProphet [71] to produce a protXML file.

For well-sampled proteomes, protein relative abundances can be estimated on the basis of spectral count values. To calculate protein abundance ratios, a normalization scheme needs to be applied such that the total spectral counts across samples are equal. Alternatively, count-based normalization methods with more complexity can be used such as DeSeq [57].

8 Metaproteomics

8.1 Principle

Metaproteomics refers to analysis of mixed proteomes. A synonym term proteogenomics is also used, to reflect the fact that detection of specific proteins in mixed cultures heavily depends on the quality of and the relatedness of the DNA sequence scaffolds. Not surprisingly, the best examples so far of large-scale metaproteomics approaches include low-complexity models, most prominently acid mine drainage communities for which high-quality metagenomic sequences are available [72]. In such low-complexity communities, proteomic coverage may be saturating when target organisms constitute 30–40% of total population. However, partial proteomes can be recovered even for minor populations (1% of total community) [72]. Metaproteomics of complex communities still present great challenges that range from difficulties with protein extraction from certain environments such as soils and sediments to the dynamic range of peptide detection, to the fragmented nature of metagenomic datasets, to noises and false positives. These and other challenges are described in detail in this excellent review [72].

The proteogenomic approach in the methylotrophic world has been so far applied to the plant phyllosphere [73] and to a marine sample [74]. The community of the plant phyllosphere has *Methylobacterium* as one of the dominant organisms, and thus a number of abundant proteins were assigned to this organism, including some of the methylotrophy-specific proteins (both MxaF-type and XoxF-type methanol dehydrogenases, formaldehyde activating enzyme, malyl-CoA lyase) [73]. In the marine sample, XoxF was found as one of the abundant proteins, even though the total population of the OM43 clade bacteria expressing this protein was estimated at only 1% [74]. In both cases, protein detection was most efficient when using the DNA scaffolds originating from the same environment. To demonstrate the importance of a perfectly matching scaffold in protein detection, we tested the same spectral dataset with inferred protein databases generated based on a perfectly matching DNA scaffold [65] and a scaffold representing closely related species [63]. In the former case, we obtained 68% inferred protein coverage, while in the latter the coverage was only 20% [65].

9 Metabolomics

9.1 Principle

While (meta)genomics provide the blueprint of metabolism and (meta)transcriptomics and (meta)proteomics provide clues about which genes and proteins are expressed in which conditions, detection of metabolites provides the ultimate evidence for what

metabolic activities take place. Obviously, metabolites that are subjects to rapid conversion, as part of a natural metabolic flux, would be difficult to detect. Certain manipulations such as a switch from one metabolic mode to another may be helpful in catching key player metabolites. Detecting and quantifying a metabolite of interest may be defining for interpretation of the genomic, transcriptomic, and proteomic data.

9.2 Sample Preparation

Since the turnover rate of the metabolites involved in central metabolism is in the range of seconds to minutes, rapid quenching for stopping the enzyme reactions and for keeping the cell membrane intact is essential for acquiring an accurate snapshot of the metabolome. Many quenching methods, including cold methanol–water mixture, cold glycerol–saline solution, cold ethanol–sodium chloride solution, or fast filtration protocol, have been developed to be used with either batch or bioreactor cultures. For example, a fast filtration protocol has been successfully applied for metabolomic analysis in methanotrophic bacteria [75]. For example, 3 mL culture samples that are collected in mid-exponential phase of growth are rapidly harvested by vacuum filtration using S-Pak™ membrane filters (0.22 μm, 47 mm) (Millipore, Billerica, MA) and washed with 3 mL of a fresh medium. The filter is then immediately transferred to a Petri dish located on the surface of a Cool Beans Chill Bucket™ (ISC Bioexpress, Kaysville, UT) at −5°C. To collect cells, the following three sequential rinse solutions are applied: 0.5 mL of 25 mM ice-cold HEPES buffer (pH 5.2), 0.5 mL of −20°C ethanol solution (75/25, v/v, ethanol/aqueous 25 mM HEPES buffer, pH 5.2), and 1.5 mL of −20°C ethanol. The resulting solution is transferred into a prechilled tube and stored in −80°C freezer until subsequent extractions.

Metabolites are extracted from microbial cells using different protocols, such as boiling ethanol solution; cold methanol or acetonitrile, either alone or in combination with an aqueous buffer; freeze–thaw cycles in methanol; and perchloric acid. A boiling ethanol solution has been applied to extract the metabolome from *Methylobacterium extorquens* AM1 and other methylotrophic bacteria [75–77]. Briefly, 1 mL of boiling HEPES-buffered ethanol solution (75/25, v/v ethanol/water, pH 5.2) is added to a given cell pellet, followed by incubation at 100°C for 3 min. The extracted cell suspension is cooled on ice for 3 min, and cell debris is removed by centrifugation at approximately 5,000 g for 5 min. The cell-free metabolite extract is centrifuged at 20,000 g for 8 min. The supernatant is transferred into a 2 mL glass vial and dried in a vacuum centrifuge (CentriVap Concentrator System, Labconco, MO, USA) to complete the drying. The dried sample, prior to analysis using LC-MS/MS, is redissolved in 100 mL purified water. The dried sample to be analyzed by GC-MS is further derivatized in two steps as follows [78]. In the first step,

the keto- groups are methoximated by the addition of 50 mL methoxyamine solution (25 mg/mL methoxyamine hydrochloride in pyridine), followed by incubation at 60°C for 30 min. In the second step, trimethylsilylation is performed by adding 50 mL TMS reagent (BSTFA/TMCS, 99:1) and heating at 60°C for 60 min. Notably, no matter which extraction method is used, introduction of internal standards (IS) is usually essential for accurate relative or absolute quantification. IS may be natural (^{12}C) or isotopically labeled (^{13}C or ^{15}N), and they can be added at different stages of sample preparation and analysis and used to monitor either the recovery after each preparation step or the performance of the subsequent LC-MS or GC-MS analyses.

9.3 LC-MS and GC-MS Instrumentation for Metabolomics

Mass spectrometry (MS)-based metabolomics, in which a separation technique such as gas chromatography (GC), capillary electrophoresis (CE), or liquid chromatography (LC) is coupled to a mass spectrometer, has been widely applied to profile metabolomes or to determine metabolite concentrations. Due to the versatile separation characteristics of LC, broader selectivity, and omission of the derivatization steps, LC-MS is often the preferred technique for metabolomic analysis. Metabolites are typically moderately to highly polar small molecules, which are often too hydrophilic to be reliably retained and separated on common reversed-phase columns (RPLC). Novel chromatographic techniques, including hydrophilic interaction liquid chromatography (HILIC), ion-pairing reverse-phase chromatography, and hybrid phase chromatography, are gaining popularity for metabolomics applications. However, some metabolites with similar physicochemical properties present challenges for LC analyses. As a result, the combination of multiple LC-based and GC-based methods for the same sample is preferred in order to increase the resolving power.

As described in recent reports, we have investigated central metabolisms of *M. extorquens* AM1, *Methylosinus trichosporium* OB3b, *and Methylomicrobium 20Z* using a combination of complementary separation techniques (RPLC, HILIC, and GC) with MS [75–77]. LC-MS/MS experiments were carried out on a Waters (Milford, MA, USA) LC-MS system consisting of a 1525 μ binary HPLC pump with a 2,777 autosampler coupled to a Quattro Micro™ API triple quadrupole mass spectrometer (Micromass, Manchester, UK). The HILIC method employing gradient elution was carried out using the previously described column (Luna NH_2, 250 × 2 mm, 5 μm; Phenomenex, Torrance, CA, USA) and the conditions as described below. Gradient elution was carried out with 20 mM ammonium acetate and 0.35% NH_4OH (28%) in water (v/v)/acetonitrile (95:5,v/v) with pH 9.7 (mobile phase A) and acetonitrile (mobile phase B). The linear gradients used were 85–0% B for 15 min, 0% B for 11 min, 0–85% B for 1 min, and 85% B for 15 min. The total run time was 42 min at

0.15 mL/min. The injection volume was 10 mL. The eluent from each LC column was directed into the ion source of the MS. Multiple reaction-monitoring (MRM) experiments were carried out as previously described [79, 80]. The dwell time for each MRM transition was 0.08 s. All peaks were integrated using MassLynx™ Applications Manager (version 4.1) software. For GC-MS, the experiments were performed using an Agilent 5973 MSD/6890 instrument (Agilent Corp, Santa Clara, CA, USA). The column was a 30 m × 0.25 mm × 0.5 μm film (RTX-5MS, Restek, Bellefonte, PA, USA). Ultra-high purity helium was used as the carrier gas in a constant flow mode of 1 mL/min, and 1 μL of a given sample was injected in splitless mode via an Agilent 7683 autosampler. The inlet temperature was set at 280°C. The temperature program began at 60°C with a hold time of 0.25 min, and then increased at 8°C/min to 280°C with a hold time of 10 min at 280°C. The ion source temperature was set to 250°C. Mass spectra were collected from m/z 40 to 500 at 3 spectra/s after a 6-min solvent delay [78].

In addition to a good retention and separation of metabolites, introduction of internal standards (IS) to the samples prior to metabolite extraction is important for reliable quantification. When complex biological extracts are injected into an electrospray ionization source, the ionization efficiency of metabolites can be suppressed or enhanced due to the presence of less volatile and co-eluting compounds [79]. By adding ^{13}C-labeled IS to the samples, especially the respective culture-derived, global ^{13}C-labeled IS, corrections can be made for the variations arising from instrumental analysis and/or sample preparation. With the introduction of a global ^{13}C-labeled IS, we were able to accurately profile more than 40 metabolites in both *M. extorquens* AM1 and *M. trichosporium* OB3b [75, 79].

9.4 Data Analysis

Many efforts have been made to develop efficient chemometric data analysis tools in our laboratory and elsewhere. Parallel factor analysis (PARAFAC) is one mathematical tool for peak deconvolution that provides accurate quantification of metabolites of interest, even in the presence of overlapping compounds. Recently, a novel PARAFAC method was reported for the analysis of nearly co-eluting ^{12}C and ^{13}C isotopically labeled metabolites on GC-MS data [76, 78, 81]. This methodology further forms the basis for dynamic ^{13}C flux analysis to determine the fate of specific metabolites in methylotrophs, via quantitative determination of ^{13}C-label uptake as a function of time [78, 81]. For example, by using ^{13}C-based metabolomics, alternative metabolic pathways for glyoxylate consumption were demonstrated [82].

More recently, $^{13}CH_4$-based analysis was used to measure fluxes through the metabolic network in *M. alcaliphilum* 20Z [77]. For the $^{13}CH_4$-labeling experiments, cells are typically

grown to mid-exponential phase. In order to quickly remove residues of ^{12}C-methane, cells can be transferred to a fresh flask, sealed, and supplemented with the same concentration of $^{13}CH_4$. Cell cultures are then harvested at defined time points (such as 0, 1, 2, 5, 10, 20, 40, and 60 min.). The mass isotopomer distributions should be corrected for the natural isotope contribution by using a matrix-based method [78] and calculated as the relative abundances of the different possible mass isotopomers of a metabolite [75]. Total ^{13}C-incorporation of a metabolite is obtained by normalizing to its total carbon number. Relative isotopic abundance (Mi) is calculated by the following equation:

$$Mi(\%) = mi \Big/ \sum_{j=0}^{n} mj,$$

where *Mi* represented the isotopic abundance for a metabolite in which ^{13}C atoms were incorporated and *n* represented the maximum number of ^{13}C atoms incorporated. Total ^{13}C-incorporation of a metabolite with *N* carbon atoms is obtained by normalizing to its total carbon number using the following equation:

$$\text{Total}^{13}\text{C-incorporation } (\%) = \sum_{i=1}^{N} (i \times Mi)/N$$

^{13}C-incorporation rate is then calculated from the initial slope of all ^{13}C-isotopologues versus time, i.e., the change of total ^{13}C-incorporation versus the times from 0 to 60 min.

We utilized multiple reaction-monitoring transitions on mass spectrometry (MRM-MS), to resolve the labeling position of ^{13}C-carbons in pyruvate [75]. When pyruvate is formed as part of the Entner–Doudoroff (ED) pathway, the initial ^{13}C incorporation should be observed in position (1). However, if pyruvate is derived as part of the glycolysis pathway, it would be expected to be labeled in position (3). MRM-MS with ^{13}C-labeled metabolomes has demonstrated that *M. alcaliphilum* 20Z, and likely other type I methanotrophs, utilizes both ED and glycolysis pathways for C1 assimilation.

10 Metabolic Modeling of C1 Metabolism

A few metabolic models focused on C1 metabolism, mostly for methanol- or methylamine-utilizing species, have been constructed [83, 84]. Access to the complete genome sequences of methylotrophic bacteria now provides new, top-down approaches for metabolic reconstruction of C1 metabolism [85, 86]. A number of genome-scale biochemical network reconstructions of methylotrophic bacteria are available in BioCyc (http://www.biocyc.org). However, these are based on automated reconstructions, which should be carefully evaluated in the light of the published data for

a specific microbe of interest (substrate consumption and biomass accumulation rates, biomass composition analysis, metabolic pathway validation via enzymatic activity, gene/protein expression, etc.) and then converted into a mathematical model that can be analyzed through constraint-based, linear programming tools such as COBRA (http://opencobra.sourceforge.net/openCOBRA/Welcome.html) or PathwayTools (bioinformatics.ai.sri.com/ptools), at a global, systems level and through nonlinear kinetic modeling at a local, more mechanistic level. In ideal situations, the reconstruction should be further validated through comparison of model predictions to phenotypic data.

11 Concluding Remarks

The omics techniques, especially in their combination, present an efficient way to approach the questions in metabolism, physiology, and ecology of specific guilds of microorganisms, such as methylotrophs. However, the specific protocols and the details of the omics-associated analyses are fluid, with the technologies changing and evolving very rapidly. The genomics field is approximately 20 years of age, and metagenomics are only a decade or so old. During this time, a succession of novel sequencing technologies have come into play, some featuring low cost with shorter sequence reads (Illumina) and some featuring longer reads at higher cost per nucleotide (PacBio). Novel technologies are continuously entering the scene, and, likely, further advances in sequencing technology are in the process of being developed. Thus, the recommendations given in this chapter are very tentative, based on the experiences of the past years. It is amazing to look back to the state of the field, as of only a few years ago, and realize that we are in a different world all together, in terms of both sequencing technologies and computational tools. Our message is, the specific physiology of a specific methylotrophs should be considered when designing systems biology-based experiments, in order to take advantage of the knowledge available on their specialized metabolism. At the same time, our minds should be open to discovering novel and unexpected pathways for metabolizing C1 compounds. The omics approaches, in their combination, should present a likely path toward such new knowledge.

Acknowledgements

This material is based upon the work supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC-0010556, and the Joint Fund of National Natural Science Foundation of China (Grant Number U1462109).

References

1. Chistoserdova L, Lidstrom ME (2013) Aerobic methylotrophic prokaryotes. In: Rosenberg E, DeLong EF, Thompson F, Lory S, Stackebrandt E (eds) The prokaryotes, 4th edn. Springer, Berlin, pp 267–285
2. Söhngen NL (1906) Űber bacteria welche methan als kőhlenstorfnarung und energiequelle gebrauchen. Zentr Bakt Parazitenk 15:513–517
3. Anthony C (1982) Biochemistry of methylotrophs. Academic, London
4. Ettwig KF, Butler MK, Le Paslier D et al (2010) Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. Nature 464:543–548
5. Haroon MF, Hu S, Shi Y et al (2013) Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. Nature 500:567–570, Erratum in: Nature (2013) 501:578
6. Knittel K, Boetius A (2009) Anaerobic oxidation of methane: progress with an unknown process. Annu Rev Microbiol 63:311–334
7. Beal EJ, House CH, Orphan VJ (2009) Manganese- and iron-dependent marine methane oxidation. Science 325:184–187
8. Marx C, Bringel F, Chistoserdova L et al (2012) Complete genome sequences of six strains of the genus *Methylobacterium*. J Bacteriol 194:4746–4748
9. Beck DAC, McTaggart TL, Setboonsarng U et al (2014) The expanded diversity of *Methylophilaceae* from Lake Washington through cultivation and genomic sequencing of novel ecotypes. PLoS One 9:e102458
10. Vuilleumier S, Chistoserdova L, Lee MC (2009) *Methylobacterium* genome sequences: a reference blueprint to investigate microbial metabolism of C1 compounds from natural and industrial sources. PLoS One 4:e5584
11. Wajid B, Serpedin E (2012) Review of general algorithmic features for genome assemblers for next generation sequencers. Genomics Proteomics Bioinformatics 10:58–73
12. Wajid B, Serpedin E (2014) Do it yourself guide to genome assembly. Brief Funct Genomics pii:elu042
13. Radajewski S, Ineson P, Parekh NR, Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. Nature 403:646–649
14. Neufeld JD, Vohra J, Dumont MG et al (2007) DNA stable-isotope probing. Nat Protoc 2:860–866
15. Neufeld JD, Wagner M, Murrell JC (2007) Who eats what, where and when? Isotope-labelling experiments are coming of age. ISME J 1:103–110
16. Chen Y, Murrell JC (2010) When metagenomics meets stable-isotope probing: progress and perspectives. Trends Microbiol 18:157–163
17. Kalyuzhnaya MG, Lapidus A, Ivanova N et al (2008) High-resolution metagenomics targets major functional types in complex microbial communities. Nat Biotechnol 26:1029–1034
18. Beck DAC, Kalyuzhnaya MG, Malfatti S et al (2013) A metagenomic insight into freshwater methane-utilizing communities and evidence for cooperation between the *Methylococcaceae* and the *Methylophilaceae*. PeerJ 1:e23
19. Kalyuhznaya MG, Martens-Habbena W, Wang T et al (2009) *Methylophilaceae* link methanol oxidation to denitrification in freshwater lake sediment as suggested by stable isotope probing and pure culture analysis. Environ Microbiol Rep 1:385–392
20. Nercessian O, Noyes E, Kalyuzhnaya MG, Lidstrom ME, Chistoserdova L (2005) Bacterial populations active in metabolism of C1 compounds in the sediment of Lake Washington, a freshwater lake. Appl Environ Microbiol 71:6885–6899
21. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7:e30087
22. Markowitz VM, Chen IM, Palaniappan K et al (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res 42:D560–D567
23. Kunin V, Hugenholtz P (2010) PyroTagger: a fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. Open J 1:1–8
24. Caporaso JG, Lauber CL, Walters WA et al (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108(Suppl 1):4516–4522
25. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R (2012) Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Microbiol Chapter 1: Unit 1E 5
26. Schloss PD, Westcott SL, Ryabin T et al (2009) Introducing mothur: open-source, platform-

independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541

27. Mitra S, Stark M, Huson DH (2011) Analysis of 16S rRNA environmental sequences using MEGAN. BMC Genomics 12(Suppl 3):S17
28. Kumar R, Eipers P, Little RB et al (2014) Getting started with microbiome analysis: sample acquisition to bioinformatics. Curr Protoc Hum Genet 82:181811–181829, Haines JL et al (eds)
29. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998
30. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–2200
31. Haas BJ, Gevers D, Earl AM (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21:494–504
32. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267
33. Oksanen J, Blanchet FG, Kindt R (2013) vegan: Community Ecology Package Version 2.0-10
34. Chistoserdova L (2011) Modularity of methylotrophy, revisited. Environ Microbiol 13:2603–2622
35. Dumont MG, Murrell JC (2005) Community-level analysis: key genes of aerobic methane oxidation. Methods Enzymol 397:413–427
36. Dumont MG (2014) Primers: functional marker genes for methylotrophs and methanotrophs. In: McGenity TJ et al (eds) Hydrocarbon and lipid microbiology protocols, Springer protocols handbooks. Springer, Berlin Heidelberg
37. Sy A, Giraud E, Jourand P et al (2001) Methylotrophic *Methylobacterium* bacteria nodulate and fix nitrogen in symbiosis with legumes. J Bacteriol 183:214–220
38. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595
39. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4:e7767
40. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
41. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools.1000 genome project data processing subgroup. Bioinformatics 25:2078–2079
42. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192
43. Oshkin I, Beck DAC, Lamb AE et al (2014) Methane fed microcosms show differential community dynamics and pinpoint specific taxa involved in communal response. ISME J doi:10.1038/ismej.2014.203
44. Hernandez ME, Beck DAC, Lidstrom ME, Chistoserdova L (2015) Oxygen availability is a major factor in determining the composition of microbial communities involved in methane oxidation. PeerJ 3:e801
45. Kerckhof FM, Courtens ENP, Geirnaert A et al (2014) Optimized cryopreservation of mixed microbial communities for conserved functionality and diversity. PLoS One 9:e99517
46. Dedysh SN, Dunfield PF (2014) Cultivation of methanotrophs. In: McGenity TJ et al (eds) Hydrocarbon and lipid microbiology protocols, Springer protocols handbooks. Springer, Berlin Heidelberg
47. Kelly DP, Ardley JK, Wood AP (2014) Cultivation of methylotrophs. In: McGenity TJ et al (eds) Hydrocarbon and lipid microbiology protocols, Springer protocols handbooks. Springer, Berlin Heidelberg
48. Kuivila KM, Murray JW, Devol AH, Lidstrom ME, Reimers CE (1988) Methane cycling in the sediments of Lake Washington. Limnol Oceanogr 33:571–581
49. Auman AJ, Stolyar S, Costello AM, Lidstrom ME (2000) Molecular characterization of methanotrophic isolates from freshwater lake sediment. Appl Environ Microbiol 66:5259–5266
50. Reim A, Luke C, Krause S, Pratscher J, Frenzel P (2012) One millimetre makes the difference: high-resolution analysis of methane-oxidizing bacteria and their specific activity at the oxic–anoxic interface in a flooded paddy soil. ISME J 6:2128–2139
51. Okubo Y, Skovran E, Guo X, Sivam D, Lidstrom ME (2007) Implementation of microarrays for *Methylobacterium extorquens* AM1. OMICS 11:325–340
52. Ojala DS, Beck DA, Kalyuzhnaya MG (2011) Genetic systems for moderately halo(alkali)

philic bacteria of the genus *Methylomicrobium*. Methods Enzymol 495:99–118

53. Matsen JB, Yang S, Stein LY, Beck D, Kalyuzhnaya MG (2013) Global molecular analyses of methane metabolism in methanotrophic Alphaproteobacterium, *Methylosinus trichosporium* OB3b. Part I: transcriptomic study. Front Microbiol 4:40
54. Anders S, Pyl PT, Huber W (2014) HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford, England). doi: 10.1093/bioinformatics/btu638
55. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628
56. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440–9445
57. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106
58. Anders S, McCarthy DJ, Chen Y et al (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 8:1765–1786
59. Gourion B, Rossignol M, Vorholt JA (2006) A proteomic study of *Methylobacterium extorquens* reveals a response regulator essential for epiphytic growth. Proc Natl Acad Sci U S A 103:13186–13191
60. Bosch G, Skovran E, Xia Q et al (2008) Comprehensive proteomics of *Methylobacterium extorquens* AM1 metabolism under single carbon and nonmethylotrophic conditions. Proteomics 8:3494–3505
61. Kappler U, Nouwens AS (2013) Metabolic adaptation and trophic strategies of soil bacteria-C1- metabolism and sulfur chemolithotrophy in *Starkeya novella*. Front Microbiol 4:304
62. Müller JE, Litsanov B, Bortfeld-Miller M et al (2014) Proteomic analysis of the thermophilic methylotroph *Bacillus methanolicus* MGA3. Proteomics 14:725–737
63. Bosch G, Wang T, Latypova E, Kalyuzhnaya MG, Hackett M, Chistoserdova L (2009) Insights into the physiology of *Methylotenera mobilis* as revealed by metagenome-based shotgun proteomic analysis. Microbiology 155:1103–1110
64. Hendrickson EL, Beck DA, Wang T, Lidstrom ME, Hackett M, Chistoserdova L (2010) Expressed genome of *Methylobacillus flagellatus* as defined through comprehensive proteomics and new insights into methylotrophy. J Bacteriol 192:4859–4867
65. Beck DA, Hendrickson EL, Vorobev A et al (2011) An integrated proteomics/transcriptomics approach points to oxygen as the main electron sink for methanol metabolism in *Methylotenera mobilis*. J Bacteriol 193:4758–4765
66. Kao WC, Chen YR, Yi EC et al (2004) Quantitative proteomic analysis of metabolic regulation by copper ions in *Methylococcus capsulatus* (Bath). J Biol Chem 279:51554–51560
67. Deutsch EW, Shteynberg D, Lam H et al (2010) Trans-Proteomic Pipeline supports and improves analysis of electron transfer dissociation data sets. Proteomics 10:1190–1195
68. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24:2534–2536
69. Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. Proteomics 13:22–24
70. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383–5392
71. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75:4646–4658
72. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: functional analysis of natural microbial consortia using community proteomics. Nat Rev Microbiol 7:196–205
73. Delmotte N, Knief C, Chaffron S (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc Natl Acad Sci U S A 106:16428–16433
74. Sowell SM, Abraham PE, Shah M (2011) Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. ISME J 5:856–865
75. Yang S, Matsen JB, Konopka M et al (2013) Global molecular analyses of methane metabolism in methanotrophic alphaproteobacterium, *Methylosinus trichosporium* OB3b, Part II: metabolomics and ^{13}C-labeling study. Front Microbiol 4:70
76. Yang S, Sadilek M, Synovec RE, Lidstrom ME (2009) Liquid chromatography-tandem quadrupole mass spectrometry and comprehensive

two-dimensional gas chromatography-time-of-flight mass spectrometry measurement of targeted metabolites of *Methylobacterium extorquens* AM1 grown on two different carbon sources. J Chromatogr A 1216:3280–3289

77. Kalyuzhnaya MG, Yang S, Rozova ON et al (2013) Highly efficient methane biocatalysis revealed in a methanotrophic bacterium. Nat Commun 4:2785
78. Yang S, Nadeau JS, Humston-Fulmer EM, Lidstrom ME, Synovec RE (2012) Gas chromatography-mass spectrometry with chemometric analysis for quantitative ^{13}C Isotope determination in bacteria: towards a platform for high-throughput dynamic flux analysis. J Chromatogr A 1240:156–164
79. Yang S, Sadilek M, Lidstrom ME (2010) Streamlined pentafluorophenylpropyl column liquid chromatography-tandem quadrupole mass spectrometry and global 13C-labeled internal standards improve performance for quantitative metabolomics in bacteria. J Chromatogr A 1217:7401–7410
80. Yang S, Synovec RE, Kalyuzhnaya MG, Lidstrom ME (2011) Development of a solid phase extraction protocol coupled with liquid chromatography mass spectrometry to analyze central carbon metabolites in lake sediment microcosms. J Sep Sci 34:3597–3605
81. Yang S, Hoggard JC, Lidstrom ME, Synovec RE (2013) Comprehensive discovery of ^{13}C labeled metabolites in the bacteria *Methylobacterium extorquens* AM1 using gas chromatography – mass spectrometry. J Chromatogr A 1317:175–185
82. Okubo Y, Yang S, Chistoserdova L, Lidstrom ME (2010) Alternative route for glyoxylate consumption during growth on two-carbon compounds by *Methylobacterium extorquens* AM1. J Bacteriol 192:1813–1823
83. Van Dien SJ, Lidstrom ME (2002) Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph *Methylobacterium extorquens* AM1, with application to reconstruction of C(3) and C(4) metabolism. Biotechnol Bioeng 78:296–312
84. Peyraud R, Schneider K, Kiefer P, Massou S, Vorholt J, Portais JP (2011) Genome-scale reconstruction and system level investigation of the metabolic network of *Methylobacterium extorquens* AM1. BMC Syst Biol 5:89
85. Santos F, Boele J, Teusink B (2011) A practical guide to genome-scale metabolic models and their analysis. Methods Enzymol 500:509–532
86. Karr JP, Sanghvi JC, Macklin DN et al (2012) A whole-cell computational model predicts phenotype from genotype. Cell 150:389–401

Protocols for Probing Genome Architecture of Regulatory Networks in Hydrocarbon and Lipid Microorganisms

Costas Bouyioukos, Mohamed Elati, and François Képès

Abstract

Genome architecture and the regulation of gene expression are expected to be interdependent. Understanding this interdependence is key to successful genome engineering. Evidence for nonrandom arrangement of genes along genomes, defined as the relative positioning of cofunctional or co-regulated genes, stems from two main approaches. Firstly, the analysis of contiguous genome segments across species has highlighted the conservation of gene order (synteny) along chromosome regions. Secondly, the study of long-range regularities along chromosomes of one given species has emphasised periodic positioning of microbial genes that are either co-regulated, evolutionarily correlated, or highly codon biased. Software tools to detect, visualise, systematically analyse and exploit gene position regularities along genomes can facilitate the studies of such nonrandom genome layouts and the inference of transcription factor binding sites and potentially guide rational genome design. Here, a computational protocol is demonstrated for the analysis and exploitation of regular patterns in a set of genomic features of interest (e.g. cofunctional or co-regulated genes, chromatin immunoprecipitation results, etc.). This case study is conducted for genes involved in hydrocarbon metabolism of a marine petroleum-degrading bacterium *Alcanivorax borkumensis*.

Keywords: Gene regulation, Genome architecture, Genome organisation, Periodicity detection, Prediction of TFBSs

1 Introduction

In trying to understand and engineer microorganisms, it proved rewarding to consider at once the threefold relation between chromosome spatial conformation, genome expression, and genome layout. Genome layout is defined here as the respective positioning of cofunctional genes. 'Cofunctional genes' refer to three, not mutually exclusive, possibilities: genes that encode proteins from the same complex or from the same metabolic pathway or genes that are co-regulated by the same regulatory factor. Indeed, individual gene transcription is modulated by sequence-specific transcription factors (TF). A TF binds to its binding site (TFBS) in the regulatory region of its target gene(s) to activate or repress its transcription. Short-range genomic similarities in the one-dimensional (1-D)

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 119–134,
DOI 10.1007/8623_2015_92, © Springer-Verlag Berlin Heidelberg 2015, Published online: 14 May 2015

positioning, known as synteny [1], reveal valuable information regarding the physiology of microorganisms. However, it has been demonstrated and is an active area of investigation that yet another level of three-dimensional (3-D) organisation of the genome is realised in terms of the long-range periodic arrangements of genomic features along genomes [2, 3]. Studies involving co-regulated [4, 5], cofunctional [6] and evolutionary correlated [7] genes have all identified sets of periodic patterns of the organisation of genes along microbial chromosomes. As these regularities in genome organisation can serve as a means for genomes to accommodate a series of physiological constraints [8, 9], the systematic detection, analysis and visualisation of such periodic patterns can elucidate regulatory mechanisms at the genomic level and provide insights for rational genome design in microorganisms.

Here, we demonstrate the use of a computational approach which detects and analyses patterns of regular organisation of the positions of genomic features of interest (e.g. genes). This approach is part of a more general schema of using modelling of genome architecture as a tool for studying and engineering regulation on a global – genomic – level. This general computational schema is called Genome REgulatory and Architecture Tools (GREAT) and is under development in our team at the institute of Systems and Synthetic Biology (iSSB). In this chapter we give a detailed account about how to use, interpret and exploit the results of the SCAN suite of tools which comprises all the analytical capacities of the GREAT schema.

The chapter is organised as follows: Section 2 provides a brief introduction to the materials and requirements to perform analyses with the GREAT:SCAN suite as well as a quick description of each tool. Section 3 is the main section where the details of each tool are delineated and the steps for a successful analysis are described further. Finally, Sect. 4 deals with troubleshooting and provides a guide for the values of the most significant parameters of the analysis.

2 Materials

GREAT:SCAN is a computational protocol for the integrated analysis of regular patterns in genomes. Its requirements are merely computational/software based. No previous programming experience is required to perform a complete GREAT:SCAN analysis, and no installation is required as, at the moment, GREAT:SCAN is available as a web tool. The required computations are performed by the calculators of the abSYNTH platform of synthetic biology at the institute of Systems and Synthetic Biology (iSSB, www.issb.genopole.fr). All the files generated during the execution (plots, tables and raw output files) can be downloaded by the user as a

single zipped file. The web interface to perform the analysis can be found at the address https://absynth.issb.genopole.fr/Bioinformatics/ by selecting the icon for GREAT. The software accepts a range of optional parameters to control the analysis steps; all the parameters are implemented in the online interface of the tool, and a technical overview is presented in Appendix 2. Every user has unlimited and free access to perform analyses with the tools, with a single requirement to complete a very simple registration process at the above-mentioned internet address.

2.1 GREAT:SCAN: patterns

GREAT:SCAN:patterns is a tool written in R and is based on concepts and algorithms previously developed by the Képès team [2, 10]. The single requirement to perform the *pattern* analysis of the software suite is the format of the input file. Every input file of the analysis should contain two columns (any additional column will be ignored by the system). The first column should contain a unique identifier (e.g. a name) of the gene or the genomic feature of interest. The second column should contain the genomic coordinate (i.e. the position in the genome) of the gene or the genomic feature of interest. A single space is sufficient as a delimiter between the two columns although the system can accept any kind of conventional delimiter (tabs, semicolons). Appendix 1 contains an example of how the input file looks like. The source of the input data is totally arbitrary and is based on the motivation and the object of study of every researcher. GREAT:SCAN:patterns has been used to identify periodicities among co-regulated and co-evolved genes as well as among sites from ChIP-Seq or transcriptomics analyses. For this reason hereafter, when we describe a generic feature of GREAT:SCAN:patterns, we will be referring to the input as the 'set of genomic features of interest' (there be positions of co-regulated genes, ChIP-Seq peaks, transcriptomics peaks or any other set of interest as long as it obeys this simple rule of having a position in the genome and a unique identifier).

2.2 GREAT:SCAN: PreCisIon

GREAT:SCAN:PreCisIon is a tool written in R and based on concepts and algorithms previously developed by the team [11, 12].

PreCisIon is a general supervised method to infer new regulatory relationships between a known TF and genes in an organism. In its current form, it requires two types of data as inputs. Firstly, each gene in the organism must be characterised by some properties (views), here two views: its promoter sequence and its chromosomal position. While the former property has been used in all TFBS prediction studies so far, the latter has been developed by our team. The tool 'retrieve-seq' of the 'Regulatory Sequence Analysis Tools' http://rsat.ulb.ac.be/rsat/ [13] was used to retrieve upstream regulatory sequences ('promoters') defined here by the DNA sequence between position −400 and −1.

Secondly, for each TF, a list of its known target genes and, if possible, of its known nontargets is needed. Such lists can be constructed from publicly available databases of experimentally characterised regulations such as RegulonDB [14].

PreCisIon splits the problem of regulatory network inference into many binary classifications from disjoint views. For each view, PreCisIon trains a binary classifier to discriminate between genes known to be regulated and nonregulated by the TF. The final step is to combine all individual classifiers that have been trained on all (two here) disjoint views. All genes known to be regulated by this TF form a class of positive examples, and no prediction is needed for them. The remaining genes are split in three subsets of roughly equal size. In turn, each subset is taken apart, and PreCisIon is trained on all the positive examples, plus all genes in the two other subsets considered as negative examples. PreCisIon is then tested on the third subset, which has not been used during training. Rotating three times over the three subsets allows PreCisIon to attribute a prediction to each unlabelled gene by using an independent model.

2.3 Bacterial Genome

Alcanivorax borkumensis is a ubiquitous marine petroleum oil-degrading bacterium with an unusual physiology specialised for alkane metabolism. Its genome sequence and its genes involved in hydrocarbon metabolism were retrieved from the published freely available genome [15] through the UCSC Archaeal Genome Browser [16].

3 Methods

3.1 GREAT:SCAN: patterns

3.1.1 Periodicity Analysis

Every GREAT:SCAN study starts with a systematic and rigorous analysis and evaluation of all the periodic patterns that can be identified in the full genome of an organism. To this end, a preprocessing step is of paramount importance, the removal of proximity effects within the set of interest. Genomic features that are close to each other can 'contaminate' the calculation of probability values (p-values) for periods, as a few genes that are in proximity to each other can give a strong periodic signal with a single gene that is sufficiently far. Furthermore, as we study long-range regularities on bacterial chromosomes, we need to remove the sequential organisation of co-regulated and cofunctional genes into operons [17]. Thus, in the first step, all operons are reduced to a single position, that of their first cistron, because it is closest to the transcriptional start point. In the second step, a set of proximal genes is replaced by a single site located at their barycentre. The proximity criterion is defined by the user by specifying the average intergenic distance for the organism under study (by default two times this average intergenic distance).

The software then executes the periodicity detection algorithm as it is described in [10] exhaustively, that is, it looks for every possible period in the set of genomic features of interest and evaluates each one independently. The periods are evaluated according to their *p*-value, after applying a correction calculation to account for multiple testing. Indeed, for relatively short periods, many periods get tested, thus increasing the chances that a significant pattern will be detected. The *p*-values are weighted to take this fact into account.

At this level the user can specify a cut-off for the *p*-values (by default the significance level of 0.05 is applied) of the periods that will get displayed. The selected *p*-values are plotted in a typical plot that is used frequently in analyses of periodic phenomena and is called the periodogram (Fig. 1). A periodogram provides a quick overview of the most significant periods in terms of *p*-values (it depicts both the initial as well as the weighted *p*-value), and the researcher can readily identify which periods (if any) are the significant ones for the set of genomic features of interest.

3.1.2 An Example from Hydrocarbon Metabolism Genes

Here, we provide a test case of our analysis by performing a full GREAT:SCAN:patterns analysis on the genes from *A. borkumensis* which are involved in alkane degradation. We manually selected all the members of the two hydrocarbon degradation systems of *A. borkumensis* from [15] and found their respective translation start sites positions along the *A. borkumensis* genome. This information is enough to generate an input file for a GREAT:SCAN analysis. The only extra information that is required is the genome length as well as the average intergenic size of *A. borkumensis* which is used by the software in order to remove genome proximity effects. The output of the software consists of a set of tables where all the relevant information for each period is collected as well as a periodogram which is depicted in Fig. 1. The GREAT:SCAN analysis of the *A. borkumensis* genes that are involved in hydrocarbon metabolism found periods which approach the full genome size of the microorganism. This indicates that the major organisational principle of the hydrocarbon metabolism genes is 1-D genomic clustering, a result that could have been speculated from the neighbouring genomic coordinates of several genes in the set. Please note that this proximal trend is detected despite the prior removal of direct neighbourhood by the algorithm. However, a significant period of around 50 kbp was detected also, a finding which can raise some interesting insights (*see* Sect. 3.1.2 and further discussion in the legend of Fig. 1).

3.1.3 Clustering In-Phase Genes

Periodically arranged genes have a specific radial position on the modulo period coordinates of each individual significant period. Visualising their modulo coordinates is of key interest for biological researchers because this view might provide insights on whether the

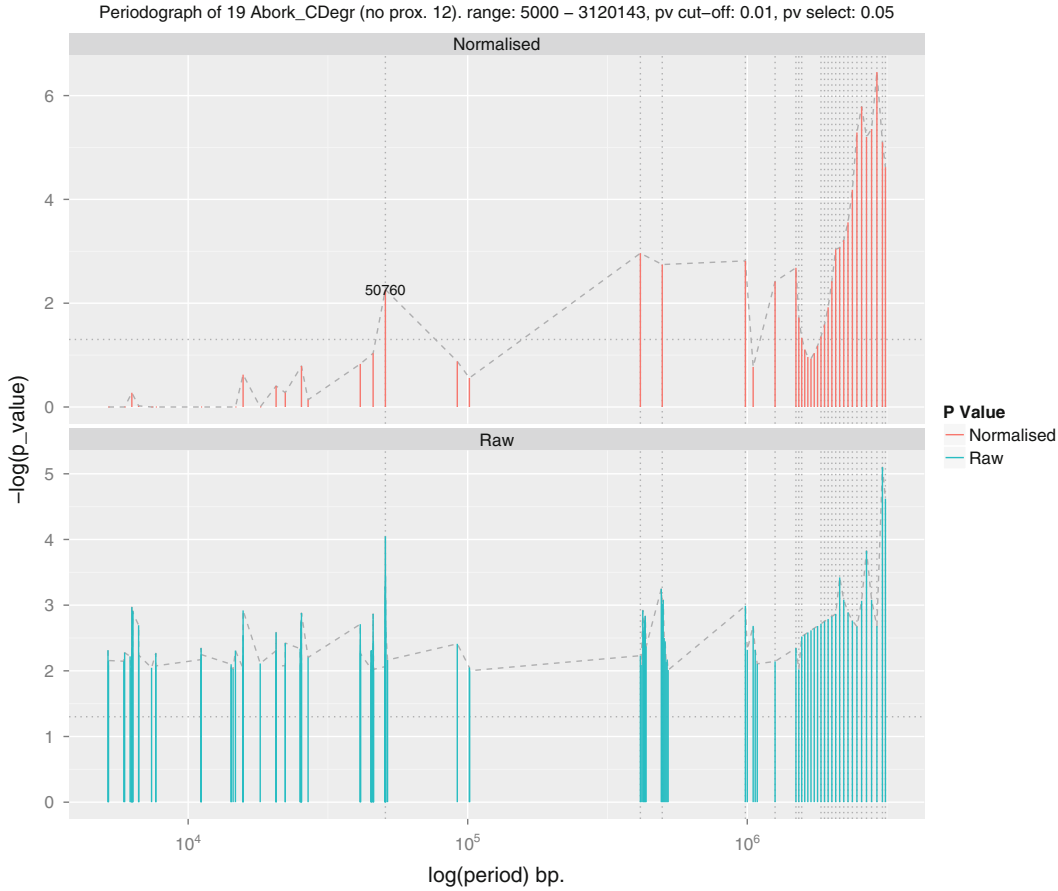


Fig. 1 Periodogram of genes involved in hydrocarbon metabolism of *A. borkumensis*. The height of the bars corresponds to the significance of the detected period (−log(*p*-value); thus if, e.g. the *p*-value equals 10^{-3}, the bar height is 3), the *dotted vertical lines* indicate highly significant periods (periods with *p*-value lower than the user specified *pvThres* parameter) and the *dashed line* connects the tips of the bars together to provide a view of regions with dense periodic signal detection. The *upper panel* depicts the same periods where the *p*-values have been normalised to correct for multiple testing and ordered by their size. The *lower panel* depicts the raw non-normalised *p*-values. Numerous significant periods are found to be close to the size of the whole *A. borkumensis* chromosome; this finding indicates that the proximal genomic arrangement of hydrocarbon metabolism genes is significant (*see* also Note 4.3 in the text). However, in the lower end of the spectrum, a few bars are also significant, which indicates a periodic pattern and suggests a potential 3-D solenoid arrangement of genes. Notably, a significant peak is detected for period 50,760 bp. A further step (Sect. 3.1.2) of the analysis with GREAT:SCAN:patterns can provide more information about that finding

set of the genes of interest can be found to be co-localised in the 3-D folding of the chromosome and take advantage of any proximity or local concentration effect for their transcriptional activity. We employ a simple density clustering approach based on an algorithm known in data sciences as DBSCAN [18]. DBSCAN is an unsupervised clustering algorithm that requires two parameters to find clusters, the minimum size of the cluster (which is set to two

genes by default) and the minimum distance between points (which is set as the ratio of the average intergenic size to the period; this ratio is normalised by a single parameter called the clustering exponent, set to 0.5 by default). This density-based clustering technique is applied to the modulo period coordinates of the genomic features of interest for all of the significant periods that have resulted from the previous periodicity analysis step (Sect. 3.1.1). For each of the significant periods, the user obtains a table of the clustered genes including their position information score as well as a unique plot for each of the significant periods that we call a clustergram. A clustergram visualises the formation of the clusters of the genomic features of interests after plotting their modulo coordinate (phase) on the *x*-axis and their phase ranking on the *y*-axis. The clustergram automatically colours the clustered genes according to the cluster they belong to. However, a viewer can also identify clusters by looking for vertical alignment of genomic features of interest in the plot (Figs. 2 and 3). Genes belonging to a cluster will appear to be perfectly aligned on a vertical line in a clustergram plot.

3.1.4 An Example from Hydrocarbon Metabolism Genes

Continuing the analysis of the genes involved in hydrocarbon metabolism of *A. borkumensis*, GREAT:SCAN:patterns computed the clustergrams of all the significant periods from the previous analysis step (*see* Sect. 3.1.1). The results from the periodogram analysis indicate that most of the significant periods are similar to the genome length thus implying a 1-D genomic proximity arrangement of the hydrocarbon metabolic genes. The clustering analysis corroborates that further, as genes are clustered for the period of 3,043,845 bp as it is demonstrated in the clustergram of Fig. 2. However, a much shorter period of 50,760 bp was also found to be significant, and the hydrocarbon-involved genes appeared to cluster well (Fig. 3), suggesting a potential 3-D clustering of hydrocarbon metabolism genes.

3.1.5 Chromosome Mapping

So far, we considered periods that span the full length of the genome. Each period analysed and studied up to this section refers to the full set of the genes (or the genomic features) of interest as it is positioned on the whole genome. However, there might be cases where only a certain chromosomal region displays periodic arrangement. This section of the protocol will provide the tools and techniques to analyse these cases too.

To address this requirement, an additional feature of the algorithm was developed: the periodicity analysis can be performed in a sliding window. We have developed a 'mapping' algorithm where a sliding window approach is scrounging the whole genome on multiple scales in order to identify periodic regions. This section (and the following) will provide the steps to perform chromosome mapping analysis and interpret the results.

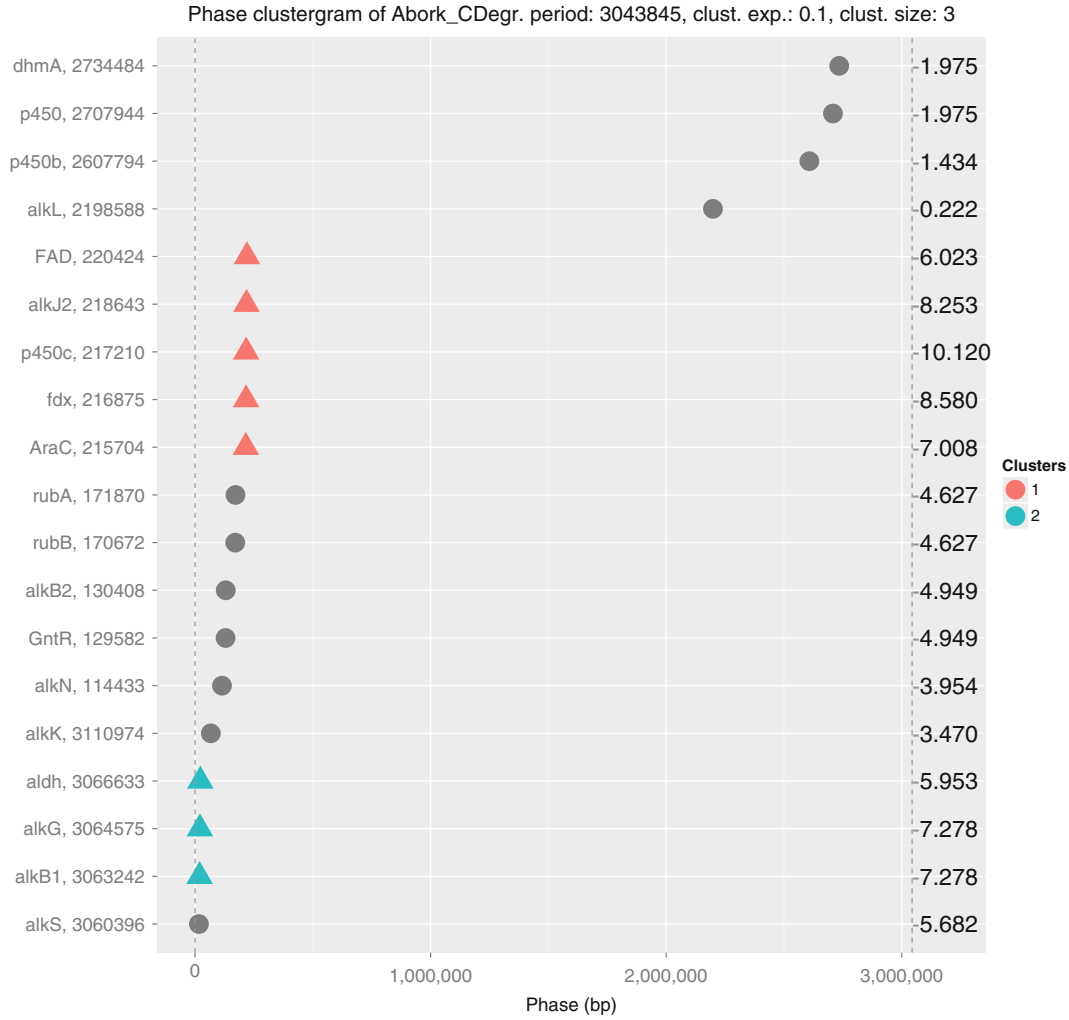


Fig. 2 Clustergram of the hydrocarbon metabolism genes of *A. borkumensis* for a period close to full genome length. The *x*-axis represents the length of one whole period, and each location corresponds to the phase (modulo coordinate) of each genomic feature of interest. Thus, any vertical quasi-alignment of the points denotes a gene cluster. The *left y-axis* shows the gene name and its genomic position; the *right y-axis* shows the positional information score of each gene (a score which corresponds to the individual contribution of each genomic feature to the clustering for this particular period). Cases like this, where the period approaches the genomic length, capture 1-D proximity, because proximity is detected by going around the *full circle* of the genome and falling back in the same neighbourhood

The period-scanning algorithm that is described in [10] and is used in section 3.1.1 to detect periods in the whole genome is adapted with a sliding window approach so that it operates in segments of the genome. The size of the window is specified by the user; however, a default value of 10,000 bp that grows incrementally to the whole genome length is used and provides the right

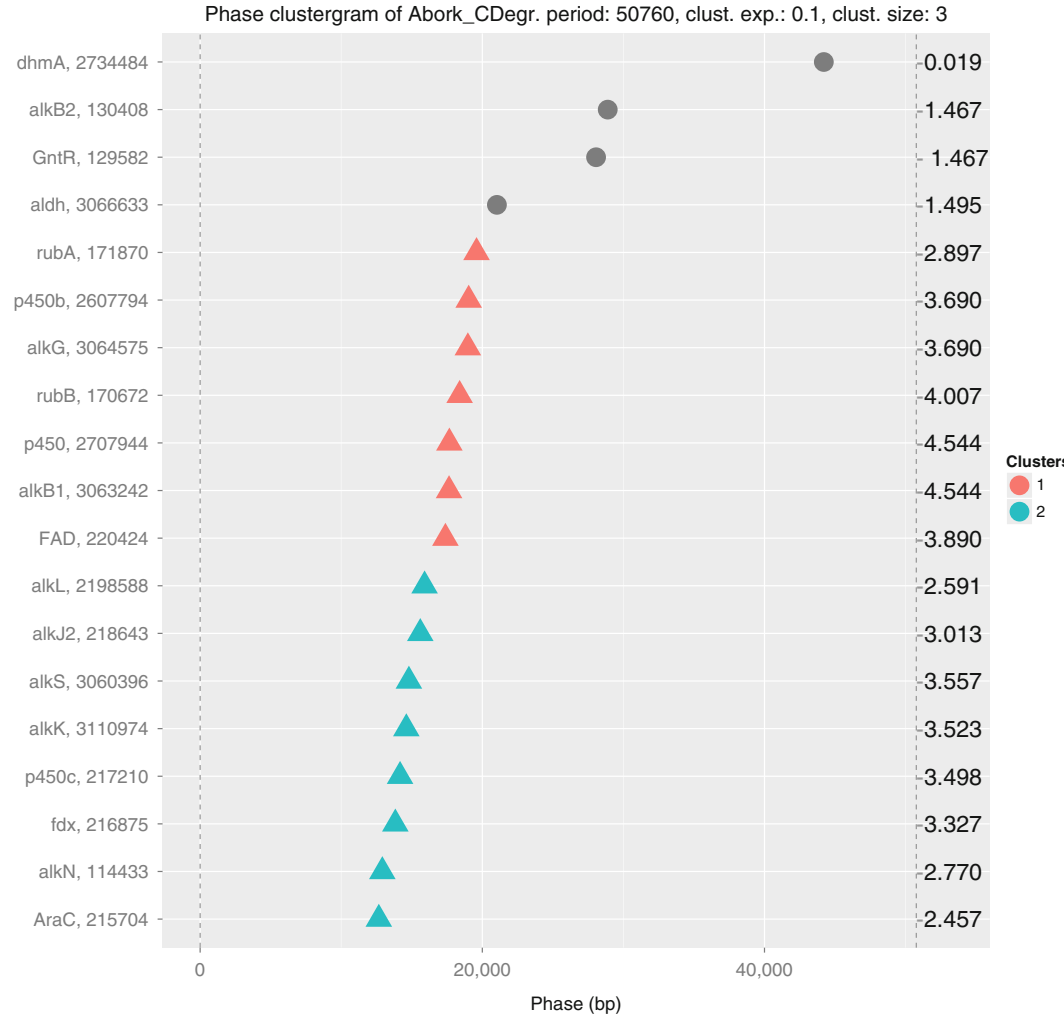


Fig. 3 Clustergram of the hydrocarbon metabolism genes of *A. borkumensis* for a period of 50,760 bp. Cases, like the one illustrated where the period is much lower than genome length, may be interpreted as 1-D periodicity, suggestive of 3-D arrangement and clustering of genes [2, 6]. Genes, or genomic features of interest, with a high position information score, are the top candidates for further investigation of potential 3-D co-localisation. The caption of Fig. 2 describes the details regarding the graph

results in any occasion. The user can also specify a *p*-value cut-off for the periods that are selected for plotting.

3.1.6 An Example from Hydrocarbon Metabolism Genes

We continue the analysis of the genes from *A. borkumensis* which play a central role in hydrocarbon metabolism by analysing their organisation in a finer scale using the 'sliding window' version of the periodicity detection algorithm. This allows the user to obtain a graph of the whole genome of the organism of interest where the detected periods on each particular segment are immediately observable together with information about the number of

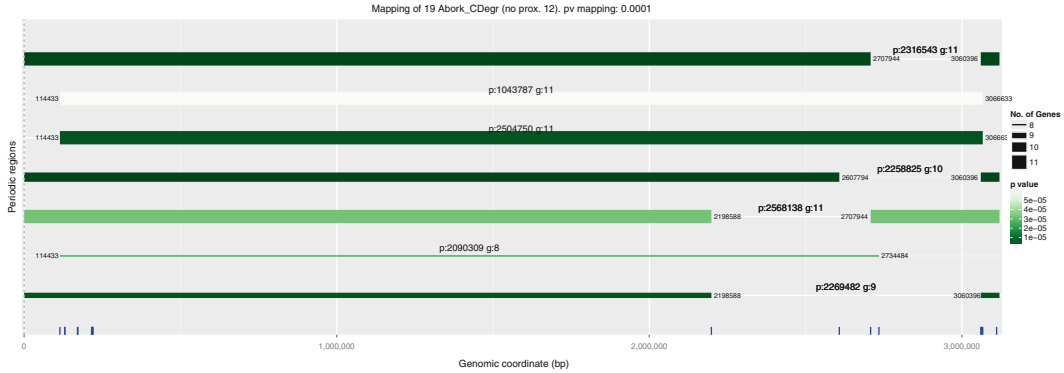


Fig. 4 The period mapping graph (or 'chromogram') of the hydrocarbon metabolism genes of *A. borkumensis*. Here, segments of the genome that contain periodic genes or genomic features of interest are depicted. The *x*-axis displays the genomic coordinates for the full genome length. The *y*-axis is used only to order the segments according to the segment size. The *thickness of the segment* denotes the number of genes that belong to this segment. The *colour code* corresponds to the *p*-value for this period. *Thickness* and *colour scales* self-adjust to the data and chosen parameters and are indicated on the *right side of the plot*. For each significant segment, its end coordinates, period value (p:) and number of genes (g:) appear in the text just above the middle of the segment. The *blue ticks* on the horizontal axis demark the genomic position of the input data. Additionally (not shown), the user can specify some genomic landmarks of interest from the parameters of the program. NOTE: all the above-mentioned plots were obtained by running the GREAT:SCAN:patterns program and using the following parameters: avgGene: 1000, clustExp: 0.1, clustSize: 3, infile: alcanivoraxHC_Metabolism.txt, length: 3120143, perRange: 5000, 3120143, pvSelect: 0.01, pvThres: 0.05, mapSelect: 0.001

involved genes, the *p*-value of each period and genomic locations of interest which can be superimposed on the graph. We call this plot a 'period mapping plot' or 'chromogram', and the result for the hydrocarbon metabolism genes is illustrated in Fig. 4.

3.2 GREAT:SCAN: PreCisIon

Current methods for the identification of cis-regulatory elements are marginally successful in their ability to discriminate between many alternative variants of the possible TFBSs. While the data on the consensus sequences for the corresponding regulatory sites are available, it often contains motifs with very low sequence conservation (like TCRNNNNNNACG, where N can be any nucleotide). Such degenerate consensus sequences lead to high false-negative and false-positive rates. The difficulty lies in the specific nature of DNA-protein interactions. Our method PreCisIon addresses this issue by taking into account both views: (a) local binding sequence readout and (b) global genome layout readout. The underlying rationale is based on the observation that co-regulated genes tend to be positioned at periodic intervals along the chromosome (*see* Sect. 3.1). The combined classifier is then obtained with an iterative weight update scheme, using a modified version of the AdaBoost algorithm. PreCisIon consistently improves methods based on consensus-binding sequence information only. This is shown by implementing a cross-validation analysis of the 20 major

transcription factors from two phylogenetically remote model organisms. For *Bacillus subtilis* and *Escherichia coli*, respectively, PreCisIon achieves on average an AUC (area under the ROC curve) of 70% and 60%, a sensitivity of 80% and 70% and a specificity of 60% and 56% [11, 12].

4 Notes

GREAT:SCAN analyses might not always detect significant periods or clusters, and might not return some (or any) plots. Even though the software tries to prevent the most common mistakes a user can make (i.e. wrong parameter choices) and suggest the most common solution, there are some cases where the plots and the results are not easily interpretable. Here, we collect a couple of these cases and give some explanations of why it happened as well as how to solve the problem.

4.1 Significant Periods Not Detected

There might be cases where significant periods will not be reported. There are two reasons why this might happen. Firstly, a genuine reason is that the input data do not contain any genomic features that are periodically arranged in the genome. Please note however that periodicity of cofunctional genes has been detected in all eubacterial phyla [6] and in baker's yeast [4]. A second reason is that the parameters for reporting the periods to plots (and tables) are very stringent, and thus none of the periods passed the thresholds. This is often the case with the region mapping algorithm. As the chromosome periodical mapping (Sect. 3.1.3) zooms on segments of the chromosome with a small portion of the whole datapoints, the *p*-value of these periods is generally much lower than the *p*-value of periods that refer to the whole genome. Therefore, the default value for the plotting of these mapped periods is much lower than the level of significance of 0.05 (set to 0.001 by default in the web server). If an analysis fails to return any periods in the chromosome mapping plot, then try to increase this threshold *p*-value.

4.2 Clusters of In-phase Features Not Detected

The clusters reported in the clustergram analysis of Sect. 3.1.2 are calculated by a local density cluster approach. The algorithm that is used is called DBSCAN, and it requires two parameters: the cluster size (by default 3) and the minimum distance between members of the cluster (specified by the clustering exponent). The clustering exponent is applied on the ratio between the period and the genome length which specifies the minimum distance parameter for clustering. The exponent ranges between 0 and 1; the closer to 1, the lower the effect of the length of the period towards clustering sensitivity is, therefore clustering becomes more sensitive for a given period. For values 0 or close to it, the minimum distance for clustering becomes the largest possible, and thus clustering

becomes less sensitive. As a rule of thumb, when no cluster of in-phase genes has been detected, it is advisable to lower the clustering exponent (the default is set to 0.5).

4.3 Period Nearly Equals Genome Length

GREAT:SCAN:patterns may return periods which equal the total genome length of the organism of interest or total genome length divided by a small integer. This was for instance the case with *A. borkumensis* (Sect. 3.1.1). Such very long periods denote significant proximity (1-D clustering) patterns. Indeed, it is known from the genome sequence of *A. borkumensis* [15] that there are several gene clusters where the hydrocarbon degradation genes are organised. This fact was evident from the periodicity analysis with GREAT:SCAN:patterns in Sect. 3.1.1, when the *patterns* procedure detects periods close to the genome length.

In sum, one interesting feature of the *pattern* algorithm is that it detects 1-D proximity and 3-D periodicity patterns in a single pass and provides *p*-values for both features that can be directly compared [10].

Acknowledgments

The authors thank the MEGA team members at iSSB for excellent discussions. This work was supported by Genopole, by the OSEO/BPI-France 'BioIntelligence' consortium and by the EU FP7 KBBE project 'ST-FLOW'.

Appendix 1: Input File Format for a GREAT:SCAN:patterns Analysis (This Example Contains the Genes from *A. borkumensis* Involved in Hydrocarbon Degradation)

```
dhmA, 2734484
alkB1, 3063242
alkB2, 130408
aldh, 3066633
alkK, 3110974
alkL, 2198588
alkN, 114433
rubB, 170672
rubA, 171870
GntR, 129582
p450, 2707944
p450b, 2607794
p450c, 217210
fdx, 216875
alkJ2, 218643
FAD, 220424
AraC, 215704
alkG, 3064575
alkS, 3060396
```

Appendix 2: Usage Message of GREAT:SCAN:patterns

```
usage: patterns.R [-h] -t [<title> [<title> ...]]
                  [-l <genome_in_bp>]
                  [-a <avgGene_in_bp>]
                  [-r [<per_bounds> [<per_bounds> ...]]]
                  [-p <pvalue_thres>]
                  [-s <pvalue_select>]
                  [-d [<set_coords> [<set_coords> ...]]]
                  [-k [<set_ticks> [<set_ticks> ...]]]
                  [-c <clust_exponent>]
                  [-z <cluster_size>]
                  [-m <pvalue_mapping>]
                  [-i [<a_uniq_ID>]] [-v <path>]
                  [-o <output_path>]
                  <file_name>
Systematically analyse, cluster and visualise results from
a complete GREAT:SCAN analysis. Full global_spectrum (-DOM
and -CIRC analysis) followed by a DBSCAN clustering to iden-
tify the in-phase genes and a solenoid_map (sliding window)
analysis and visualisation of the spread of all the possible
periods.

positional arguments:
  <file_name>     The input file consisting of two columns of
                  data formatted like this: <entity_ID>
                  <entity_position>


optional arguments:
  -h, -help       show this help message and exit
  -t [<title> [<title> ...]], -title [<title> [<title> ...]]

                  A substring to specify a title for the
                  experiment
                  (default: None)

  -l <genome_in_bp>, -chrom_length <genome_in_bp>

                  The length in bp of the organism
                  chromosome
                  (default: 4639675)

  -a <avgGene_in_bp>, -avg_gene <avgGene_in_bp>

                  The average gene length of the organism
                  genes
                  (default: 1000)

  -r [<per_bounds> [<per_bounds> ...]], -period_range
[<per_bounds> [<per_bounds> ...]]

                  The range (min. - max.) within which peri-
                  ods will be considered for further analy-
                  sis (default: 5000)
```

-p <pvalue_thres>, -pvalue_thres <pvalue_thres>

The unweighted *p*-value threshold for considering a period for further analysis (default: 0.05)

-s <pvalue_select>, -pvalue_select <pvalue_select>

The weighted *p*-value threshold for selecting which periods will be printed (default: 0.05)

-d [<set_coords> [<set_coords> ...]], -plot_coords [<set_coords> [<set_coords>...]]

Specifies a set of genomic coordinates to be printed as significant genome marks in the mapping plot (the *E.coli* macrodomains are defaults:
[46396, 603158, 1206296, 2180612, 2876552, 3758076])

-k [<set_ticks> [<set_ticks> ...]], -plot_ticks [<set_ticks> [<set_ticks>...]]

Specifies a set of axis ticks to be printed as indicators of genome marks in the mapping plot (must be equal size with the coordinates).
(default: ['ori', 'right', 'R/ter', 'ter/L', 'left', 'ori'])

-c <clust_exponent>, -clust_exp <clust_exponent>

The clustering exponent. Assigns the minimum distance d between two points to be members of the same cluster. Specifies the exponent of the ratio between the length of the period and chromosome length (p/L). (default: 0.5)

-z <cluster_size>, -clust_size <cluster_size>

The minimum number of members for a group to be considered as a cluster (DBSCAN parameter)
(default: 2)

-m <pvalue_mapping>, -pvalue_map <pvalue_mapping>

The weighted *p*-value threshold for selecting which sliding window periods will be plotted (default: 0.001)

-i [<a_uniq_ID>], -uniq_ID [<a_uniq_ID>]

The unique ID for the generation of the results folder. (default: patternAnalysis_ xxxx_xx_xx)

```
-v <path>, -pv <path>
                The path to the 'pv' fit parameters file.
                (default:   <installation_of_cmdline_
                programs>)

-o <output_path>, -output_path <output_path>
                The absolute path for a directory (exist-
                ing one including the trailing slash '/')
                where the output will be kept, or omit for
                the current working directory. (just the
                path, the directory name itself is con-
                trolled by the -i option).
                (default: <current_working_dir>)
```

References

1. Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15 (5):583–589
2. Dorman CJ (2013) Genome architecture and global gene regulation in bacteria: making progress towards a unified model? Nat Rev Microbiol 11:349–355
3. Képès F, Vaillant C (2003) Transcription-based solenoidal model of chromosomes. ComPlexUs 1:171–180
4. Képès F (2004) Periodic transcriptional organization of the *E.coli* genome. J Mol Biol 340:957–964
5. Képès F (2003) Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. J Mol Biol 329:859–865
6. Junier I, Hérisson J, Képès F (2012) Genomic organization of evolutionarily correlated genes in bacteria: limits and strategies. J Mol Biol 419:369–386
7. Wright MA, Kharchenko P, Church GM, Segré D (2007) Chromosomal periodicity of evolutionarily conserved gene pairs. Proc Natl Acad Sci U S A 104:10559–10564
8. Ma Q, Ying X (2013) Global genomic arrangement of bacterial genes is closely tied with the total transcriptional efficiency. Genomics Proteomics Bioinformatics 11:66–71
9. Porcar M, Danchin A, de Lorenzo V (2014) Confidence, tolerance, and allowance in biological engineering: the nuts and bolts of living things. Bioessays 37:95–102
10. Junier I, Hérisson J, Képès F (2010) Periodic pattern detection in sparse boolean sequences. Algorithms Mol Biol 5:31
11. Elati M, Fekih R, Nicolle R, Junier I, Herisson J, Képès F (2011) Boosting binding sites prediction using gene's positions. In: Algorithms in bioinformatics (WABI'11), LNCS – 6833, pp 92–103
12. Elati M, Nicolle R, Junier I, Fernández D, Fekih R, Font J, Képès F (2013) PreCisIon: PREdiction of CIS-regulatory elements improved by gene's positION. Nucleic Acids Res 41(3):1406–1415
13. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc 3(10):1578–1588
14. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res 41:D203–D213
15. Schneiker S, Martins dos Santos VAP, Bartels D, Bekel T, Brecht M, Buhrmester J, Chernikova TN, Denaro R, Ferrer M, Gertler C, Goesmann A, Golyshina OV, Kaminski F, Khachane AN, Lang S, Linke B, McHardy AC, Meyer F, Nechitaylo T, Pühler A, Regenhardt D, Rupp O, Sabirova JS, Selbitschka W, Yakimov MM, Timmis KN, Vorhölter F-J, Weidner S, Kaiser O, Golyshin PN (2006) Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensis*. Nat Biotechnol 24:997–1004

16. Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM (2006) The UCSC archaeal genome browser. Nucleic Acids Res 34:D407–D410
17. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in *Escherichia coli*: genomic analyses and predictions. Proc Natl Acad Sci U S A 97:6652–6657
18. Ester M, Kriegel H, Sander J, Xu X (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han J, Fayyad UM (eds) Proceedings of the second international conference on knowledge discovery and data mining (KDD-96), Portland. AAAI, pp 226–231

A Practical Protocol for Integration of Transcriptomics Data into Genome-Scale Metabolic Reconstructions

Juan Nogales and Lucía Agudo

Abstract

In recent years, an avalanche of data in the form of the so-called omics has been generated in biological sciences. Nevertheless, the effective use of this huge volume of data is challenging from a purely mathematical and statistical point of view, and integrative approaches are becoming a necessity. Genome-scale metabolic models offer an unprecedented chance to integrate and contextualise, in the correct biological context, this large amount of omics data being generated. This chapter provides a step-by-step protocol for the integration of transcriptomics data in genome-scale metabolic models by constructing condition-specific bacterial models. Subsequently, they are used to increase the accuracy of the in silico predictions in terms of metabolic flux prediction and for the better contextualisation of the transcriptomics data in the correct biological context. Two models environmental bacterial such as *Pseudomonas putida* KT2440 and *Synechocystis* sp. PCC 8063 and their corresponding GEMs are used here for such proposes.

Keywords Constraint-based reconstruction and analysis, Genome-scale model, GIMME, Omics integration, *Pseudomonas putida*, Synechocystis

1 Introduction

One of the prime biological questions still remaining is the complete deciphering of the complex genotype-phenotype relationships. Metabolic phenotype, which is understood as the metabolic fluxes displayed by a living organism against environmental perturbations, remains elusive to fully understand despite the knowledge of genotype. This is because of the complex network of interactions including signalling, regulatory and metabolic networks taking place between the biological components present in any given organism [1–3] (Fig. 1). With the advent of omics technologies, it is now possible to quantitatively monitor the levels of cellular components (e.g. nucleic acids, proteins and metabolites) enabling the identification of patterns in genotype expression and molecular interactions and, as a consequence, shortening the existing gap between genotype and phenotype. In recent years an avalanche of transcriptomics,

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 135–152,
DOI 10.1007/8623_2015_98, © Springer-Verlag Berlin Heidelberg 2015, Published online: 16 June 2015

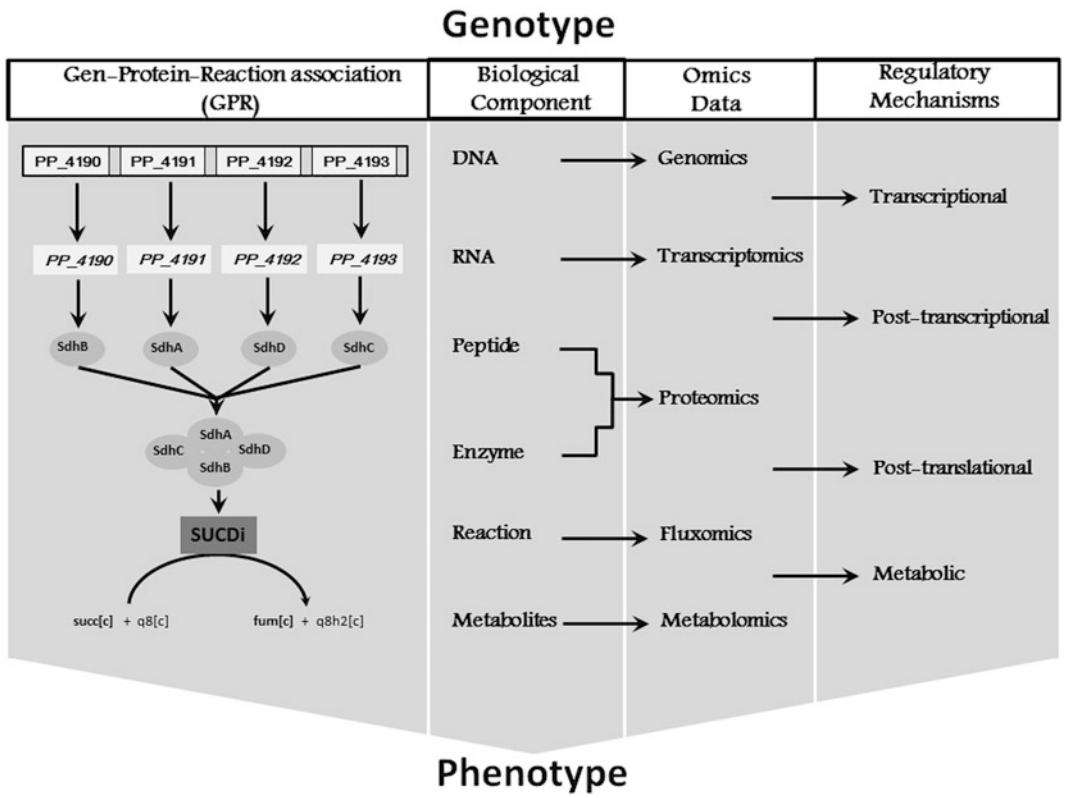


Fig. 1 Graphical representation of the gene/protein-reaction association (GPR) present in any GEM and its suitability for multi-omics data integration. Genomics, transcriptomics, proteomics, fluxomics and metabolomics data can be integrated easily in GEMs through the GPR. Additionally, the construction of condition-specific models by using different kinds of omics offers the possibility to identify new insights in the regulatory mechanism connecting both omics data

proteomics, metabolomics and fluxomics data are being collected during biological experiments. However, the correct managing and effective use of this myriad of data are challenging and integrative approaches are needed. Many so-called inference-based methods have emerged recently in order to extract biological meaning to omics datasets [4]. These methods rely on mathematical and statistical analysis of the data in order to construct biological networks, and subsequently they find common patterns in genotype expression and interactions between biological components under perturbations. These methods have proven to be highly useful; however, they are limited, at some extent, by the incomplete biological network constructed solely based on omics data which are often incomplete, hampering the contextualisation of the data in the correct biological context [4, 5].

An attractive and complementary approach comes from the possibility to integrate these data into accurate biological networks constructed from biochemical and genetic data, e.g. genome-scale

metabolic reconstructions (GEMREs). Such reconstructions contain detailed information on the target organism including the exact reaction stoichiometry, the reaction reversibility; the relationships between genes, proteins and reactions (GPR) (Fig. 1) as well as the biochemical and physiological data available [6, 7]. Thereby they are structured and species-specific knowledge bases which provide the suitable framework for omics data contextualisation [8, 9]. In addition, these reconstructions are amenable to transformation on mathematical models, genome-scale models (GEMs), which enable computation of the phenotype, in terms of metabolic fluxes, by making certain assumptions [10]. The constraint-based reconstruction and analysis (COBRA) approach [11, 12] is based on the application of those constraints imposed by the genotype and environment perturbations (e.g. nutritional conditions), and it describes the set of feasible metabolic states (*see* Nogales [9] in this protocol book series for more details about GEMs reconstruction and analysis). Because GEMs are constructed containing all the metabolic genes and reactions in a given organism, it is assumed that any reaction is active in any condition. However, many of the genes encoded in any genome are only active under particular conditions. Thus, additional regulatory constraints such as those derived from omics data can be further imposed through the GPR (Fig. 1), in order to reduce the feasible metabolic states to that corresponding to the specific environmental perturbation where the omics data were collected. Hence, this approach increases the accuracy of the model predictions. In addition, it allows the mechanistic interpretation of the data in the correct biological context, and because different kinds of omics data can be used to construct condition-specific models (e.g. transcriptomics and proteomics), the comparison of these models constructed based on different sources of omics data can provide new insights into the regulatory mechanism connecting both omics data (e.g. post-transcriptional mechanisms, Fig. 1) [8, 13, 14].

From those omics now being collected, metabolic fluxes [15] can be integrated into GEMs in a straightforward manner by constraining the maximum and minimum fluxes of the reactions in the model to those experimentally measured. However, the wide application of such approach remains challenging because the large facilities required and the reduced number of fluxes and conditions that can be experimentally measured [16], thus making its use not applicable at genome scale. Contrary, the advancements in high-throughput methods for the quantification of nucleic acids and proteins have emphasised their use as more suitable omics technologies for constraining the solution space in GEMs [8]. In particular, the use of transcriptomics data has become popular not only due to its low cost but also because changes in mRNA concentration can be determined with great accuracy due to its higher coverage when compared with proteomics. Therefore, it is not

surprising that several constraint-based methods for integration of transcriptomics data into GEMs have been developed recently (*see* for review [8, 13, 14, 17, 18]). Roughly, the current methods can be classified into two main categories [8]: (a) the switch approach, based on constraining the fluxes of the reactions whose encoding gene is under a user-given threshold expression level (e.g. GIMME [19]), and (b) the valve approach, which constrains the fluxes of the reactions in the network based on relative gene expression (e.g. PROM [20]). Recently, a very detailed comparison of these methods revealed that none of the methods published so far outperforms the others and it was concluded that there is no universal method for addressing accurately all of the metabolic scenarios tested [18]. Beyond this shortcoming is the fact that gene/protein expression levels do not necessarily reflect flux levels. Therefore, some of the methods yielded reasonable predictions under certain conditions, thus suggesting that the method of choice is not a trivial issue and that a given biological problem should be addressed through more than one method (*see* Machado and Herrgård [18] for more detailed discussion about the methods performance). For practical reasons, we will focus here in switch-based methods since they only require a limited number of gene expression datasets as input, often only a single dataset, and because they have been extensively used. Specifically, Gene Inactivity Moderated by Metabolism and Expression algorithm, GIMME [19], will be used here (Fig. 2). GIMME is a switch-based method which, although initially developed to create tissue-specific human cells models, can be easily adapted to create condition-specific bacterial models. By minimising the use of reactions encoded by genes with expression levels under a given threshold, GIMME finds a flux distribution consistent with a biological objective [19]. In others words, GIMME constructs condition-specific models by removing from the original GEM those reactions whose encoding genes are considered unexpressed in the omics dataset. The resulting model has to satisfy an appropriate objective function (e.g. biomass production); otherwise, the method restores the model functionality by adding the minimal set of the reactions that satisfy the objective (Fig. 2). GIMME has been extensively used for the study of multicellular metabolic processes in human tissues [21] and for the construction of host-pathogen [22] and disease-specific models [23], among other applications. In addition, GIMME has the advantage that it can be used to input both transcriptomics and proteomics datasets allowing the extension of this protocol to proteomics data, and it is already implemented in the popular COBRA Toolbox [12]. Therefore, it does not require additional implementation making it a user-friendly method. In this protocol we use GIMME to construct condition-specific models from two environmental bacteria such as *Pseudomonas putida* KT2440 and *Synechocystis* sp. PCC 6803 by using transcriptomics data. We further use these models

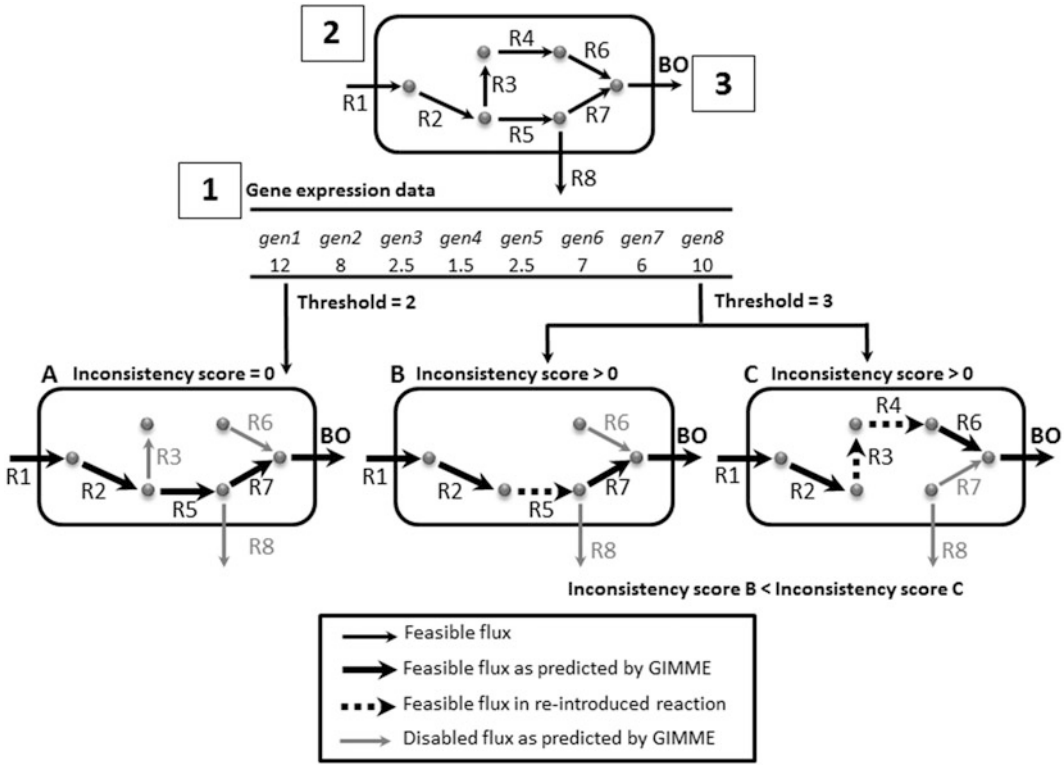


Fig. 2 Schematic representation of GIMME. GIMME is a switch-based method that uses a set of gene expression data (*1*), a GEM (*2*) and a biological objective (BO, *3*) which is assumed to be achieved by the cell under the conditions where the gene expression data were collected. GIMME is applied in two steps. First FBA is applied on the original GEM to find a flux distribution optimising the imposed BO. Then, by comparing the mRNA transcripts level in the experimental gene expression dataset with a user-imposed threshold, the method identifies inactive reactions whose encoding genes have an mRNA transcript level lower than the given threshold. Subsequently the inactive reactions are removed from the GEM and FBA is applied again on the reduced GEM. If the new model is able to achieve the assumed BO, the condition-specific model is ready. Otherwise, GIMME reintroduces in the model sets of inactive reactions by minimising the deviation from the expression data. For this, an inconsistency score is calculated for each reaction by multiplying the distance between the mRNA transcript level and the threshold value per the optimal flux required to achieve the imposed BO through the target reaction. The reaction or set of reactions allowing model functionality while minimising the inconsistency score are back into the system. The toy networks constructed to illustrate how GIMME works is composed of eight reactions R1–8, eight genes *gene1–8* and six metabolites (*2*). The gene expression values in a given condition are expressed in arbitrary units (*1*). The biological objective assumed is also indicated (*3*). Transcriptomics-based condition-specific models are constructed by using two different thresholds. When a threshold value of 2 is used, only the *gene4* is assumed to be unexpressed and subsequently its associated reaction is removed from the network (*A*). The assumed BO is achieved meaning that the expression data are consistent with the assumed functionality. As consequence the condition-specific model has an inconsistency score $= 0$. The feasible and disabled fluxes by GIMME are shown as *bold* and *grey arrows*, respectively. When a threshold value of 3 is used, three genes are considered unexpressed, *gene3–5*. This higher threshold results in a condition-specific model unable to satisfy the imposed BO which indicates that the gene expression data are only partially consistent with the assumed functionality. Subsequently, GIMME calculates the inconsistency score for each inactive reaction (e.g. R3–5) and returns a functional model minimising the overall inconsistency score. By assuming the same maximum flux value for all of the reactions in the network and in order to achieve the BO from the two potential models that can be constructed (*b* and *c*), GIMME returns to the model B since the inconsistency score is lower. The flux through the inactive reactions reintroduced in the model is shown as *dotted arrows*

to improve the flux distribution prediction in *P. putida* as well as to enhance our understanding of the photosynthetic robustness in *Synechocystis*.

2 Materials

1. Equipment.
   - 1.1 A personal computer capable of running Matlab.
   - 1.2 Matlab, version 7.0 or above (The MathWorks Inc., Natick, MA, http://www.mathworks.es/). Matlab is a numerical computing environment.
   - 1.3 COBRA Toolbox version 2.0 or above. The latest version can be downloaded from http://opencobra.sourceforge.net/openCOBRA/Welcome.html. The COBRA Toolbox is a set of MATLAB scripts for constraint-based modelling which run within the MATLAB environment.
   - 1.4 libSBML programming library 4.0.1 or above (http://sbml.org/Software/libSBML).
   - 1.5 SBMLToolbox version 3.1.1 or above for MATLAB to allow reading and writing models in SBML format (http://www.sbml.org).
   - 1.6 A linear programming (LP) solver. The COBRA Toolbox supports several open-access and commercial solvers. For the present case we used:
     - 1.6.1 GLPK (http://www.gnu.org/software/glpk) provided by the COBRA Toolbox.
     - 1.6.2 Gurobi: a free licence is available upon request at http://www.gurobi.com/.
2. Equipment Setup.
   - 2.1 Install Matlab.
   - 2.2 Install libSBML, the SBML Toolbox and selected solvers according to their specific instructions.
   - 2.3 Unpack the COBRA 2.0 archive.
   - 2.4 Initiate Matlab and navigate to the COBRA Toolbox directory.
   - 2.5 Save the path.
   - 2.6 Initialise the COBRA Toolbox by typing *initCobraToolbox*.

   Detailed information about the material, software and equipment setup can be found in Schellenberger et al. [12] and on the COBRA Toolbox website http://opencobra.sourceforge.net/openCOBRA/Welcome.html.

3. Source of GEMs. Any GEM in sbml or mat format and accounting with GPR association can be used following this protocol. An updated list of the current GEM can be found at the SBRG web page (http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms). Here we used the metabolic model of *P. putida* KT2440 (*i*JN746) [24] which can be downloaded from BIGG database (http://bigg.ucsd.edu/), and an updated model of *Synechocystis* (*i*JN678_v1.1) [25] which contain corrections in mass and charge balancing with respect to the original model [26] and can be downloaded from (http://emciblab.com/lab-members/juan-nogales.html).

3 Methods

3.1 Construction and Analysis of a Glucose Minimal Media GEM of P. putida (see Note 1)

1. Getting the GEM in proper format compatible with the COBRA Toolbox:
   1.1 Download the model of *P. putida* *i*JN746 in sbml format from the BIGG database (*see* step 2.3) and save it as P.putida_ iJN746.sbml. Place it in your Matlab work path. The basic tutorial of Matlab and detailed examples can be found in http://www.mathworks.com/support/?s_tid=gn_supp.
   1.2 Import the Model to Matlab by typing the following command:

   ```
   model=readCbModel('Pputida_iJN746.xml').
   ```

   1.3 Save the model as mat file:

   ```
   save Pputida_iJN746
   ```

2. Constructing the gene expression structure array:
   2.1 Download the gene expression data from the original source (*see* **Note 2**).
   2.2 Extract the genes from the GEM:

   ```
   Genes=model.genes
   ```

   A vector containing the genes present in the reconstruction will be created.
   2.3 Map the expression data for the list of genes in the GEM from the original gene expression data (*see* **Note 3**). Note that transcripts of 5,254 genes were detected in the current experimental study [27] and that only 746 genes are included in *i*JN746.
   2.4 From the gene expression levels (RPKM) of the genes present in *i*JN746, compute the value of the first quartile. This value will be used as a threshold (*see* **Note 4**).

2.5 Creating the presence/absence calls vector such that those genes with expression value upper and lower to the threshold will receive a value of 1 and 0, respectively.

2.6 Modify the gene ID from the original GEM to a solely numerical ID. For instance, change the original ID PP_0059 to 0059 (*see* **Note 5**).

```
model.genes = strrep(model.genes,'PP_','');
model.grRules = strrep(model.grRules,'PP_','');
```

2.7 Construct the empty gene expression structure array containing two fields (e.g. Locus and Data):

```
ExpressionData=struct('Locus',[],'Data',[])
```

2.8 Repeat step 3.1.2.2 and fill up the field "Locus" with the list of genes and the field "Data" with the presence/absence call vector in binary form (step 3.1.2.5)

3. Constructing the condition-specific model:

3.1 Constrain the exchange reactions in the model as possible, based on the experimental evidences (*see* **Note 6**). In this case, glucose uptake rate from Chavarría [28] was used.

```
model=changeRxnBounds(model,'EX_glc(e)',
-4.79,'l');
```

3.2 Add any other known constraints (*see* **Note 6**). In this case, it was assumed that the peripheral metabolism of glucose to gluconate occurs exclusively through glucose kinase [28]; thus, the flux through glucose dehydrogenase was constrained to 0.

```
model=changeRxnBounds(model,'GLCDpp',0,'b');
```

3.3 Save the constrained model.

```
save ModelGlc_Base
```

3.4 Create the condition-specific model by using the createTissueSpecificModel function from the COBRA Toolbox. This function uses as mandatory inputs a GEM in mat format (in this case ModelGlc_Base) and an expression data structure (step 3.1.2.7) (*see* **Note 7**).

```
[ModelGlc,RxnsGlc] = createTissueSpecificModel
(ModelGlc_Base,ExpressionData,[],[],
[],'GIMME');
```

The function returns two array structures.

ModelGlc contains the autotrophic-specific GEM including the regulatory constraints imposed by the transcriptomics data.

RxnsGlc contains the functional categorising of the reactions from the original GEM based on the transcriptomics data and the threshold applied (*see* **Note 8**).

3.5 Save the condition-specific model as *ModelGlc*

```
Save Pputida_Glc
```

4. Analysing the accuracy of the condition-specific model (*see* **Note 9**):

   We tested the accuracy of the glucose-specific GEM of *P. putida* by comparing the flux distribution predictions through central metabolism against those reported experimentally.

   4.1 Find the flux distribution maximising biomass production for the glucose-specific model (*see* **Note 10**).

      4.1.1 Maximise biomass production using FBA.

      ```
      Sol_ModelGlc = optimizeCbModel(ModelGlc,
      'max','one');
      ```

      4.1.2 Print the metabolic fluxes yielding such biomass maximisation.

      ```
      FluxModelGlc = printfluxVector(ModelGlc,
      Sol_ModelGlc.x,false,false)
      ```

   4.2 Find the flux distribution maximising biomass production for *i*JN746 using glucose as the only carbon source.

      4.2.1 Maximise biomass production using FBA.

      ```
      Sol_Base_Glc = optimizeCbModel
      (ModelGlc_Base,'max','one')
      ```

      4.2.2 Print the metabolic fluxes yielding such biomass maximisation.

      ```
      FluxGlc_Base = printfluxVector(SolGlcBase,
      Sol_Base_Glc.x,false,false)
      ```

   4.3 Compare flux distribution prediction of the glucose-specific model and *i*JN746 against experimentally reported flux distribution (Fig. 3).

3.2 Construction and Analysis of an Autotrophic-Specific Metabolic Model of Synechocystis *Under Optimal Light Conditions*

1. Getting the GEM in proper format compatible with the COBRA Toolbox:

   1.1 Download the model of *Synechocystis* in sbml format (*see* step 2.3). Place it in your Matlab path.

   1.2 Import the model to Matlab by typing the following command:

   ```
   model = readCbModel('iJN678_v1.1.xml')
   ```

   1.3 Save the model as a mat file:

   ```
   save iJN678_v1.1
   ```

2. Construct the gene expression structure array.

   2.1 Download the gene expression data from the original source (Note 1 [29]).

   2.2 Extract the gene from the GEM; type:

   ```
   Genes = model.genes
   ```

a

Reaction	In vivo	ModelGlc_Base	ModelGlc
GLCabcpp	100	100	100
HEX1	100	100	100
G6PDH2	100	90.38	90.41
PGL	100	90.38	90.41
PGI	0	9.59	9.55
GND	4	0	0.00
EDD	96	90.38	90.41
EDA	96	90.38	90.41
PYK	86	52.45	74.05
PC	47	0	0
PDH	129	97.73	119.83
ME1	18	0	0
FUM	104	76.76	77.52
MDH	87	77.35	99.52

b

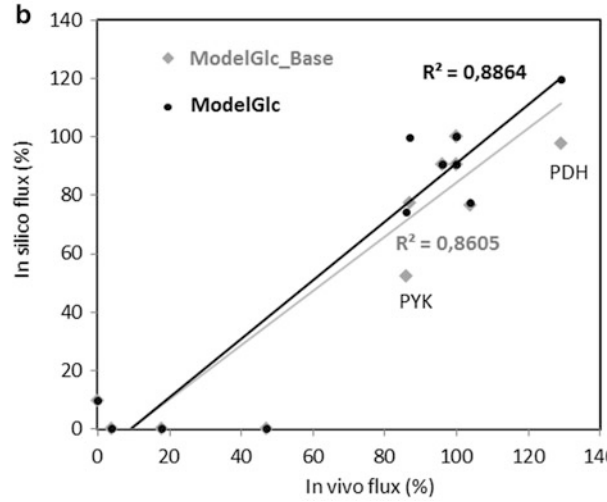


Fig. 3 Flux distribution predictions compared with experimental data. The flux predictions through the central metabolism from the original model (ModelGlc_Base) and the condition-specific model (ModelGlc) against experimental data are show in panel **A**. The flux values are shown in % and were normalised to the glucose uptake rate, which was 4.79 mmol.gDW^{-1}.h^{-1}. The correlation coefficients of the in silico predictions against experimental value are shown in panel **B**. Note that the imposition of regulatory constraints in the form of gene expression data improves significantly the in silico predictions. The condition-specific model predicts accurately higher carbon flux being funnelled to TCA driven by the pyruvate kinase (PYK) and pyruvate dehydrogenase (PDH). *GLCabcpp* D-glucose transport, *HEX1* hexokinase, *G6PDH2* glucose-6-phosphate dehydrogenase, *PGL* phosphogluconolactonase, *PGI* glucose-6-phosphate isomerase, *GND* phosphogluconate dehydrogenase, *EDD* 6-phosphogluconate dehydratase, *EDA* 2-dehydro-3-deoxy-phosphogluconate aldolase, *PYK* pyruvate kinase, *PC* pyruvate carboxylase, *PDH* pyruvate dehydrogenase, *ME1* malic enzyme, *FUM* fumarase, *MDH* malate dehydrogenase

A vector containing the genes present in the reconstruction will be created.

2.3 Map the expression data for the list of genes in the GEM from the original gene expression data. Note that transcripts of 3,106 genes were detected in the current experimental study [29] and that only 678 genes are included in *i*JN678.

2.4 From the gene expression levels (RPKM) of the genes present in *i*JN678, compute the value of the first quartile. This value will be used as a threshold (*see* **Note 4**).

2.5 Creating the presence/absence calls vector such that those genes with expression value upper and lower to the threshold will receive a value of 1 and 0, respectively.

2.6 Modify the gene ID from the original GEM to a solely numerical ID. For instance, change the original ID **sll**1682 to **1**1682 (*see* **Note 5**).

```
model.genes = strrep(model.genes,'sll','1');
model.grRules = strrep(model.grRules,'sll','1');

model.genes = strrep(model.genes,'slr','2');
```

```
model.grRules = strrep(model.grRules,'slr','2');
model.genes = strrep(model.genes,'sml','3');
model.grRules = strrep(model.grRules,'sml','3');
model.genes = strrep(model.genes,'smr','4');
model.grRules = strrep(model.grRules,'smr','4');
model.genes = strrep(model.genes,'ssl','5');
model.grRules = strrep(model.grRules,'ssl','5');
model.genes = strrep(model.genes,'ssr','6');
model.grRules = strrep(model.grRules,'ssr','6');
```

2.7 Construct the empty gene expression structure array containing two fields (e.g. Locus and Data) by typing:

```
ExpressionData = struct('Locus',[],'Data',[])
```

2.8 Repeat step 3.2.2.2 and fill up the field "Locus" with the list of genes and the field "Data" with the presence/absence call vector in binary form (step 3.2.2.5)

3. Constructing the condition-specific model:

3.1 Constrain the exchange reactions in the model in order to simulate the environmental conditions corresponding to the transcriptomics data (*see* **Note 6**).

In this case, autotrophic conditions under optimal light conditions will be used [26, 29].

Inorganic carbon is supplied in form of hco3 [26].

```
model = changeRxnBounds(model, 'EX_hco3(e)',
-3.70, 'l');
model=changeRxnBounds(model,'EX_glc(e)',0,'l');
```

Unconstrained light uptake is allowed.

```
model = changeRxnBounds(model, 'EX_photon(e)',
-100, 'l');
```

Biomass formation under autotrophic condition is selected as BO.

```
model = changeObjective(model, 'Ec_biomass_
SynAuto');
```

3.2 Add any other physiological constraint susceptible to be included, such as known metabolic fluxes. In this case, it was assumed no exchange of CO_2 under these conditions.

```
model=changeRxnBounds(model,'EX_co2(e)',0,'l');
```

3.3 Save the constrained model.

```
save ModelAuto_Base
```

3.4 Create the condition-specific model by using the createTissueSpecificModel function from the COBRA Toolbox. This function uses mandatory inputs, a GEM in

mat format (in this ModelAuto_Base) and an expression data structure (step 3.2.2.7) (*see* **Note 7**).

```
[ModelAuto,RxnsAuto] = createTissueSpecificModel
(ModelAuto_Base,ExpressionData,[],[],
[],'GIMME');
```

The function returns two array structures.

ModelAuto contains the autotrophic-specific GEM including the regulatory constraints imposed by the transcriptomics data.

RxnsAuto contains the functional categorising of the reactions from the original GEM based on the transcriptomics data and the threshold applied.

3.5 Save the condition-specific model as *ModelAuto*.

4. Analysing of the condition-specific model (*see* **Notes 8** and **9**): The autotrophic-specific metabolic model of *Synechocystis* was analysed in terms of the flux prediction accuracy and for better understanding of the metabolic states feasible under optimal light conditions.

4.1 Analysis of the flux distribution accuracy

4.1.1 Find the flux distribution maximising biomass production under autotrophic conditions under optimal light conditions using the autotrophic-specific model (ModelAuto).

Maximise biomass production using FBA.

```
Sol_ModelAuto = optimizeCbModel
(ModelAuto,'max','one');
```

Print the metabolic fluxes yielding such biomass maximisation.

```
FluxModelAuto = printfluxVector(ModelAuto,
Sol_ ModelAuto .x,false,false)
```

4.1.2 Find the flux distribution maximising biomass production for *i*JN678 using autotrophic conditions under optimal light conditions (**see** steps 3.2.3.1 and 3.2.3.2).

Maximise biomass production using FBA.

```
Sol_ModelAuto_Base = optimizeCbModel
(ModelAuto_Base,'max','one')
```

Print the metabolic fluxes yielding such biomass maximisation.

```
FluxModelAuto_Base = printfluxVector(Mode-
lAuto_Base, Sol_ModelAuto_Base.x,false,
false)
```

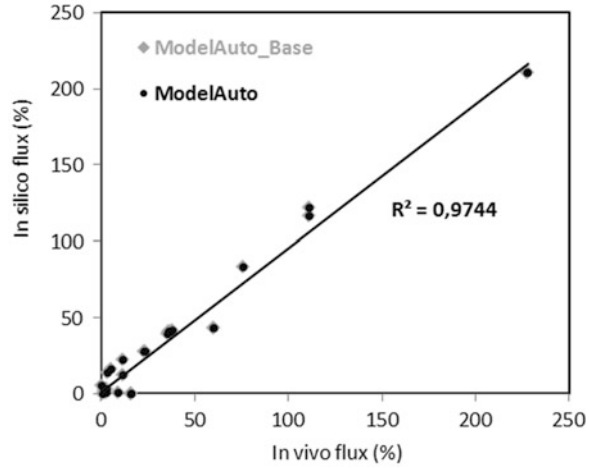


Fig. 4 Flux distribution predictions compared with experimental data. The construction of an autotrophic-specific model of *Synechocystis* under optimal light conditions does not increase in this case the accuracy of the flux predictions which was extremely high even in the original model

4.1.3 Compare flux distribution prediction of the autotrophic-specific model (ModelAuto) and *i*JN678 (ModelAuto_Base) against experimentally reported flux distribution (Fig. 4).

4.2 Exploring of the feasible metabolic states using random sampling

4.2.1 Sample the solution space in the autotrophic-specific metabolic model (*see* **Note 11**).

```
[sampleStructModelAuto, mixedFracAuto] =
gpSampler(ModelAuto, 4000,[], 57600, [])
```

4.2.2 Sample the solution space in *i*JN678 under autotrophic conditions (*see* **Note 11**).

```
[sampleStructModelAuto_Base, mixedFracAu-
to_Base] = gpSampler(ModelAuto_Base, 4000,
[], 57600, [])
```

4.2.3 Compare the metabolic states feasible through the linear electron flow pathway in the model of *Synechocystis* under autotrophic conditions with and without the regulatory constraints imposed by transcriptomics data.

Select the reactions of interest.

```
rxnList = {'PSII','CBFCu','PSI','F-
NOR','RBPC', 'ATPSu'};
```

Plot the feasible flux distribution for each reaction in both models.

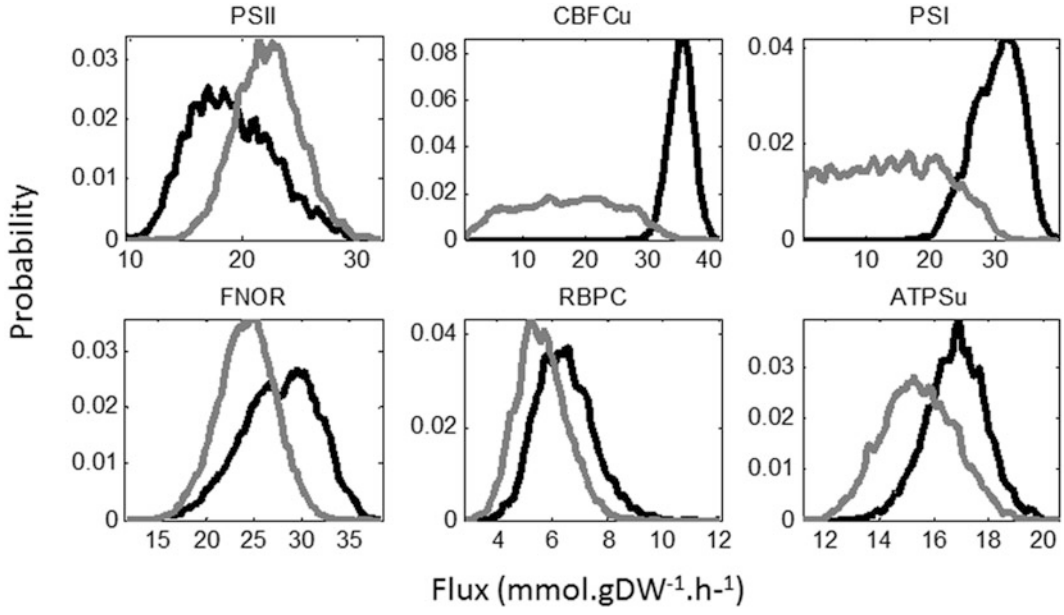


Fig. 5 Impact of the introduction of regulatory constraints on the flux distribution through the linear electron flow photosynthetic pathway in *Synechocystis*. The linear electron flow (LEF) pathway in *Synechocystis* is composed by the photosystem II (PSII), cytochrome Cytb6f (CBFCu), photosystem I (PSI) and ferredoxin-NADP reductase (FNOR). The LEF is assisted by multiple alternate electron flow pathways as a function of the light conditions in optimal photosynthesis performance. As a consequence, LEF adopts multiple flux distributions in response to the AEF pathway(s) activated in a given light condition [26]. This behaviour is consistent with the large range of flux distributions predicted through the LEF using sampling in the original model (ModelAuto_Base, *grey*) which accounts for the complete set of metabolic states feasible in *Synechocystis*. However, the imposition of the regulatory constraints and the construction of an autotrophic-specific model of *Synechocystis* under optimal light conditions (ModelAuto, *black*) reduce significantly the number of metabolic states possible (e.g. remove those corresponding to nonoptimal light conditions). As a result, the range of flux distributions through the LEF narrows significantly. Therefore, despite the condition-specific model unable to increase the accuracy regarding flux predictions, it was able to delimit the metabolic states feasible to that specifically corresponding to the environmental condition analysed, showing the robustness of this approach

```
plotSampleHist(rxnList, {sampleStructMo-
delAuto.points,sampleStructModelAuto_-
Base.points}, {ModelAuto,
ModelAuto_Base},[1000],[3, 3]);
```

See Fig. 5.

4 Notes

1. The GIMME algorithm already implemented in the COBRA Toolbox [12] requires as inputs a GEM in mat format and an expression structure array containing: (a) a vector for gene IDs and (b) a presence/absence call vector for each gene in the reconstruction in binary form where 1 and 0 indicate presence

and absence, respectively. Be sure that the two vectors from the expression structure array have the same dimension.

2. Multiple sources of transcriptomics data can be scrutinised from primary literature to general gene expression repository databases such as Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and KEGG Expression Database (http://www.genome.jp/kegg/expression/), up to species-specific gene expression databases. For a recent evaluation of these sources, *see* Rung and Brazma [30]. Although less developed, similarly public proteomics repository databases such as Proteomics IDEntifications (PRIDE) http://www.ebi.ac.uk/pride/archive/ and Multi-Omics Profilling Expression Database (MOPED) https://www.proteinspire.org/MOPED/mopedviews/proteinExpressionDatabase.jsf are available and contain large datasets of already published protein expression studies. Here, absolute gene expression transcriptomics data of *P. putida* growing in glucose minimal medium [27] and *Synechocystis* under autotrophic conditions [29] were used.
3. Due to the large number of databases and associated gene/protein IDs, the conversion of the gene/protein identifiers to that included in the proper GEM is a key step. For most of the current GEMs, the gene IDs included in the GPR are unique and correspond to the gene IDs used in the popular metabolic databases KEGG. However, for many expression datasets, alternative gene IDs can be used, such as those used in databases like EntrezGene, RefSeq, UniGene, etc. In the same way, many expression datasets include platform-specific gene IDs such as those from Affymetrix, Agilent, etc. For these cases, several freely available ID conversion tools are available and can be used for mapping gene IDs on GEMs (*see*, e.g., http://hum-molgen.org/NewsGen/08-2009/000020.html).
4. Since the bacterial gene expression is continuous, there are no well-established rules to consider a given gene significantly expressed or not. Thus, the user-imposed threshold for considering whether a gene is significantly expressed is one of the most sensitive parameters when using GIMME. A popular method for selecting systematically the threshold is to consider that the gene is not significantly expressed if its expression level is under the first quartile [18]. In addition, it is recommended to compute the threshold value only taking into account metabolic genes (these are those included in the reconstruction) since non-metabolic genes such as those encoding for tRNAs or ribosomal proteins present very high levels of expression [31]. The construction of condition-specific models by using different thresholds is therefore desirable previously to the final analysis.

5. The function implemented in the COBRA Toolbox is design specifically for constructing tissue-specific models using as input the human GEM which contains numerical genes ID. Therefore, the conversion to numerical IDs for genes is required to bypass this technical limitation without modifying the function in the COBRA Toolbox.
6. The model needs to be constrained by using experimental nutrients uptake rates and/or by-product secretion rates measured under the environmental condition to simulate. In addition, any other physiological constraints susceptible to be included, such as known metabolic fluxes, should be included as well.
7. The method implemented in the COBRA Toolbox offers the possibility to include additional inputs. Here default conditions have been used including the achievement of 90% of the biomass produced by the original model.
8. The RxnsGlc structure contains the categorising of the reaction from the original model based on the transcriptomics data of their coding genes. ExpressedRxns are those predicted to be expressed, UnExpressedRxns are those predicted to be unexpressed and are removed from the model, and Upregulated are those added back into the model in order to achieve the BO while minimising the inconsistence score.
9. The imposition of regulatory constraints based on the transcriptomics data excludes from the new model those metabolic state consequences of the predicted unexpressed genes, reducing the solution space. Thus, it is expected that the accuracy of the condition-specific model increases significantly with respect to the original one. These condition-specific models can be analysed using the large array of COBRA methods currently available [11].
10. A detailed description of flux balance analysis (FBA) can be found in Orth et al. [10]. A practical tutorial is also available in this book of protocols [9].
11. The random sampling method implemented in the COBRA Toolbox uses the hit-and-run algorithm [32], and it computes the probabilistic flux value for each single reaction by obtaining points uniformly distributed in the region of allowed solutions [33]. The gpSampler function returns two outputs. The sampleStructModelAuto contains the sampling structure including the flux distribution points allowed for each reaction in the network. The mixedFracAuto is the statistical index which regards the quality of the sampling analysis. A value of 0.5 means that the solution space has been sampled uniformly. A total of 4,000 points were used in this protocol (it is recommendable to use at least double number of points than reactions in the model). In addition, the sampling was run for 57,600 s. The rest of the variables were used as defaults.

Acknowledgements

The authors would like to thank C. Herencias for testing the protocol and valuable discussion.

The research leading to these results has received funding from the Ministry of Economy and Competitiveness of Spain Grant BIO2012-39501, European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 311815 (SYNPOL project, http://www.synpol.org/) and European Union's H2020 ERAIB LigBio project (PCIN-2014-113).

References

1. Ray JCJ, Tabor JJ, Igoshin OA (2011) Non-transcriptional regulatory processes shape transcriptional network dynamics. Nat Rev Microbiol 9(11):817–828
2. Chubukov V, Gerosa L, Kochanowski K, Sauer U (2014) Coordination of microbial metabolism. Nat Rev Microbiol 12(5):327–340
3. Kochanowski K, Sauer U, Chubukov V (2013) Somewhat in control – the role of transcription in regulating microbial metabolic fluxes. Curr Opin Biotechnol 24(6):987–993
4. De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nat Rev Microbiol 8(10):717–729
5. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G (2010) Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci U S A 107(14):6286–6291
6. Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5(1): 93–121
7. Feist A, Herrgard M, Thiele I, Reed J, Palsson B (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7:129–143
8. Hyduke DR, Lewis NE, Palsson BO (2013) Analysis of omics data with genome-scale models of metabolism. Mol Biosyst 9(2):167–174
9. Nogales J (2014) A practical protocol for genome-scale metabolic reconstructions. Humana Press, New York, pp 1–25
10. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28 (3):245–248
11. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10(4):291–305
12. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S et al (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc 6(9):1290–1307
13. Blazier AS, Papin JA (2012) Integration of expression data in genome-scale metabolic network reconstructions. Front Physiol 3
14. Kim MK, Lun DS (2014) Methods for integration of transcriptomic data in genome-scale metabolic models. Comput Struct Biotechnol J 11(18):59–65
15. Wiechert W (2001) 13C metabolic flux analysis. Metab Eng 3(3):195–206
16. Crown SB, Antoniewicz MR (2013) Publishing 13C metabolic flux analysis studies: a review and future perspectives. Metab Eng 20:42–48
17. Saha R, Chowdhury A, Maranas CD (2014) Recent advances in the reconstruction of metabolic models and integration of omics data. Curr Opin Biotechnol 29:39–45
18. Machado D, Herrgård M (2014) Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol 10(4), e1003580
19. Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol 4(5), e1000082
20. Chandrasekaran S, Price ND (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 107 (41):17845–17850
21. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, Cheng JK, Patel N, Yee

A, Lewis RA, Eils R et al (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nat Biotechnol 28(12):1279–1285

22. Bordbar A, Lewis NE, Schellenberger J, Palsson BØ, Jamshidi N (2010) Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. Mol Syst Biol 6:422
23. Chang RL, Xie L, Xie L, Bourne PE, Palsson BØ (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. PLoS Comput Biol 6(9), e1000938
24. Nogales J, Palsson B, Thiele I (2008) A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. BMC Syst Biol 2(1):79
25. Gudmundsson S, Nogales J (2015) Cyanobacteria as photosynthetic biocatalysts: a systems biology perspective. Mol Biosyst 11 (1):60–70
26. Nogales J, Gudmundsson S, Knight EM, Palsson BO, Thiele I (2012) Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. Proc Natl Acad Sci U S A 109(7):2678–2683
27. Kim J, Oliveros JC, Nikel PI, de Lorenzo V, Silva-Rocha R (2013) Transcriptomic fingerprinting of *Pseudomonas putida* under alternative physiological regimes. Environ Microbiol Rep 5(6):883–891
28. Chavarría M, Kleijn RJ, Sauer U, Pflüger-Grau K, de Lorenzo V (2012) Regulatory tasks of the phosphoenolpyruvate-phosphotransferase system of *Pseudomonas putida* in central carbon metabolism. MBio 3(2):e00028-12
29. Anfelt J, Hallström B, Nielsen J, Uhlén M, Hudson EP (2013) Using transcriptomics to improve butanol tolerance of *Synechocystis* sp. strain PCC 6803. Appl Environ Microbiol 79 (23):7419–7427
30. Rung J, Brazma A (2013) Reuse of public genome-wide gene expression data. Nat Rev Genet 14(2):89–99
31. Zur H, Ruppin E, Shlomi T (2010) iMAT: an integrative metabolic analysis tool. Bioinformatics 26(24):3140–3142
32. Lovász L (1999) Hit-and-run mixes fast. Math Program 86(3):443–461
33. Schellenberger J, Palsson BØ (2009) Use of randomized sampling for analysis of metabolic networks. J Biol Chem 284(9):5457–5461

Computer-Guided Metabolic Engineering

M.A. Valderrama-Gomez, S.G. Wagner, and A. Kremling

Abstract

Computational methods and tools are nowadays widely applied for rational Metabolic Engineering approaches. However, what is still missing are clear advices on the right order of the application of these tools. The availability of genomic information for a large number of cellular systems especially requires the use of computers to store, analyze, and process knowledge of single enzymes, metabolic pathways, and cellular networks. The trend of integrating measured quantities for the metabolome, the transcriptome, and the proteome into mathematical models, combined with methods for the rational design of cellular networks, has led to the research field Systems Metabolic Engineering, a field that extends and amplifies the classical field of Metabolic Engineering. This chapter describes mathematical and computational approaches on the cellular and the process levels. In the Material section, modeling approaches and methods for model analysis are introduced, and the current state of the art is reviewed. In the Method section, we propose a protocol for efficiently combining various approaches for the optimal production of desired biotechnological products.

Keywords: Constraint-based modelling, Dynamic flux balance analysis, Flux balance analysis, In silico strain optimization, Metabolic Engineering, Metabolic models, Stoichiometric analysis, Succinate production, Systems Metabolic Engineering, Theoretical yields

1 Introduction

Computational methods and tools are nowadays widely applied for rational Metabolic Engineering approaches. The optimization of hydrocarbon and lipid production or degradation is one concrete example for the application of this tool and has already been applied successfully by a number of research groups [1–4]. Usually, a large amount of biological data is necessary for Metabolic Engineering. For example, the availability of genomic information for a large number of cellular systems especially requires the use of computers to store, analyze, and process knowledge of single enzymes, metabolic pathways, and cellular networks. The trend of integrating measured quantities for the metabolome, the transcriptome, and the proteome into mathematical models, combined with methods for the rational design of cellular networks, has led to the research

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 153–184,
DOI 10.1007/8623_2015_118, © Springer-Verlag Berlin Heidelberg 2015, Published online: 26 July 2015

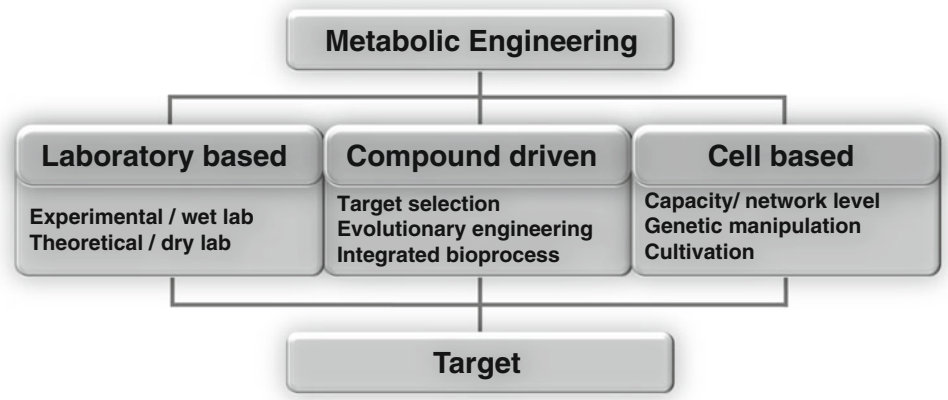


Fig. 1 Views describing the field of Metabolic Engineering. Each view tackles the engineering problem of the efficient target production using different tools and focuses. This chapter describes theoretical tools used in the laboratory-based view

field Systems Metabolic Engineering, a field that extends and amplifies the classical field of Metabolic Engineering. In recent years different views have become popular that describe the field (Fig. 1).

All views address a different question but have a common aim, namely, the efficient production of a target. Depending on the applied tools, methods, and researcher expertise, the cell-based view, the compound-driven view, or the laboratory-based view predominate. The cell-based view starts by exploring the capabilities of the cellular systems and modifies enzymes, pathway elements, or network elements to optimize the system [5]. Furthermore, the ease of genetic manipulation and cultivation of the cells is the driving force. The compound-based view [6] starts with the target component and asks how the component can be synthesized. The laboratory-based view distinguishes between experimental and theoretical approaches.

In the theoretical laboratory-based view, computational tools must perform diverse tasks to support the optimization of cellular systems with respect to the production of desired compounds. Three main tasks can be identified: search for information in genomic and metabolic databases, description and integration of experimental data in mathematical models, and application of optimization strategies to improve single enzymes, to design pathways and networks, and to construct new cellular circuits.

For the first of these tasks, databases like KEGG [7], EcoCyc [8] or Brenda [9] provide information about compounds, reactions, and networks for various cellular systems. Moreover, kinetic information, that is, information on the temporal behavior of enzymes, can also be found. In general, the information is very detailed, ranging from the chemical structure of compounds and promoter and ribosome sequences to pathway information. In this way, databases support all strategies for modification of cellular systems. The setup and the analysis of mathematical models are further pillars in Systems Metabolic Engineering. Such models are

helpful in two ways: first, they integrate what we know of a system in terms of mathematical equations. Since these equations are based on physical and chemical laws, models are used to check the consistency of the knowledge and thereby allow researchers to detect missing or incorrect items. Second, quantitative models, that is, models that are validated with quantitative experimental data, have potential for prediction. The chance to make predictions about conditions that were not used for model validation opens possibilities for model-based modifications such as optimization of cellular properties. Optimization itself plays the most important role in Systems Metabolic Engineering and is used not only on the cellular level but also on the process level.

When optimizing the metabolic system with respect to the production of a target, there are two possible cases. Figure 2 shows that either the target is already inherently produced by the host cell (case A), or a noninherent pathway has to be inserted (case B). In case A, a metabolite is often produced only at a low specific rate r, while the organism is growing at a high growth rate (Fig. 2a, left). The aim is to construct a strain with a higher specific production rate. Caused by a reorganization of the available resources, a lower growth rate results (Fig. 2a, right). If the target is not produced by the strain inherently (Fig. 2b, left), heterologous DNA information should be introduced into the strain. When new enzymes are expressed and the desired product is built, it is expected that the growth rate also decreases (Fig. 2b, right).

This chapter describes mathematical and computational approaches on the cellular and the process level. In the Material section, modeling approaches and methods for model analysis are introduced, and the current state of the art is reviewed. In the Method section, we propose a protocol for efficiently combining various approaches for the optimal production of desired biotechnological products.

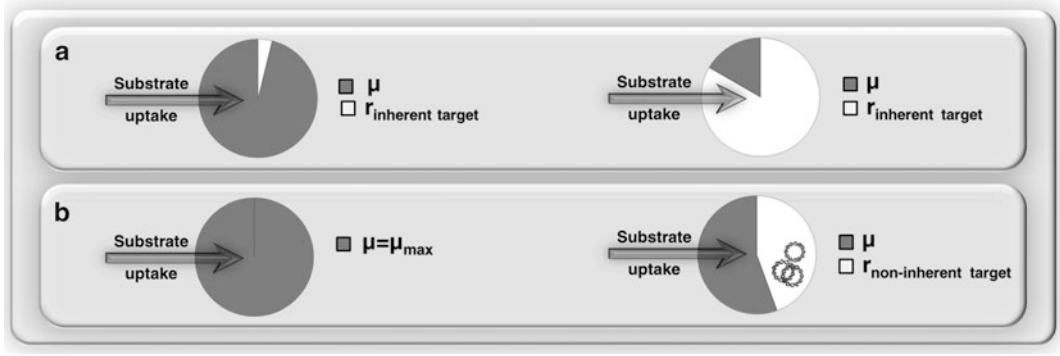


Fig. 2 Resource usage in two different cases. The *circles* represent the available cellular resources and how they are used in different situations: (**a**) the target is already produced inherently by the cell (*left*). Optimization of this metabolic system results in a rearrangement of the resources (*right*). (**b**) If the target is not produced naturally (*left*), a noninherent pathway has to be introduced. This also causes a reorganization of the cellular metabolism. Here, the plasmids represent the heterologous DNA introduced into the host cell (*right*)

2 Materials

Commonly used materials for computer-guided Metabolic Engineering include metabolic reconstructions, model equations for the cellular reaction network and for the bioreactor system, experimental data, and software. The last category included solvers, computing environments, and/or programming languages.

2.1 Metabolic Reconstruction

A genome-scale metabolic reconstruction is a mathematical representation of the metabolism of a living cell. It is typically made up of the stoichiometry of all known reactions that take place inside an organism and the enzymes and genes associated with that reaction. Additionally, a reaction accounting for biomass generation [Eq. (1)] which is based on the biomass composition of that microorganism and an estimation for growth- (GAM) and nongrowth-associated energy requirements (NGAM) are also important components of the metabolic reconstruction. The concept of GAM and NGAM for the description of the energetics of bacterial cell growth was first mathematically formalized by Pirt [10]. GAM accounts for the energy needed to synthetize macromolecules (DNA, RNA, lipids, etc.) necessary for cell growth, while NGAM refers to the energy consumed for functions other than production of new cellular material.

$$\text{Protein} + \text{RNA} + \text{DNA} + \text{lipid} + \text{lipopolysaccaride} + \cdots + \text{energy} \xrightarrow{\mu} \text{biomass}. \tag{1}$$

High-quality genome-scale metabolic reconstructions for many industrially important microorganisms are freely available in public repositories (http://sbrg.ucsd.edu/Downloads). The methods for building those reconstructions are also well established [11]. Curated metabolic models can be used in constraint-based modeling approaches for the estimation of metabolic capabilities of the cell, hypothesis testing and generation, and Metabolic Engineering [12].

The scope and coverage of the metabolic reconstructions can vary substantially. Table 1 shows the evolution of the genome-scale metabolic reconstruction of *Escherichia coli (E. coli)* over the last decade [13, 14].

During this period of time, many new reactions have been introduced, and some others have been updated based on newly available biochemical knowledge. The selection of a specific metabolic reconstruction depends on the aim of the simulations to be performed. However, it is highly recommended to start with a small version (core model) of the metabolic network of interest. This allows a better understanding of the methods used while keeping an overview of the results obtained.

Table 1
Evolution of the metabolic reconstruction of *E. coli*. Exchange reactions are related to reactions that permit mass exchange between the cell and culture media

	E. coli core	iJR904 [15]	iAF1260 [16]	iJO1366 [17]
Included genes	137	904	1260	1366
Reactions	95	931	2077	2251
Exchange reactions	20	143	298	329
Metabolites	72	761	1039	1136

2.2 Model Equations

In this section, mathematical equations describing the dynamics of metabolite concentration inside the cell and the reactor are discussed. The idealized case of perfect mixing, in which no spatial concentration gradients are considered, is assumed for the reactor and the cell. The mass balance for the intracellular metabolites is formulated for an average cell, which is assumed to be representative of the whole cell population.

2.2.1 Intracellular Reaction Networks and Constraint-Based Modeling

Biochemical reactions taking place in a cell can be generically written as:

$$\begin{aligned} |\gamma_{A1}|\, A + |\gamma_{B1}|\, B &\xrightarrow{r_1} |\gamma_{C1}|\, C + |\gamma_{D1}|\, D \\ |\gamma_{A2}|\, A + |\gamma_{X2}|\, X &\xrightarrow{r2} |\gamma_{F2}|\, F + |\gamma_{G2}|\, G. \\ &\vdots \end{aligned} \tag{2}$$

γ_{ij} are stoichiometric coefficients and A, B, X, D, F, and G represent network components. The corresponding mass balance for each intracellular component reads:

$$\frac{dC_A}{dt} = \gamma_{A1} \cdot r_1 + \gamma_{A2} \cdot r_2 \cdots - \mu \cdot C_A \,, \tag{3}$$

where C_A is the concentration of component A in the cell [mmol gDW^{-1}] and r_i represents the reaction rate of the reaction i [mmol gDW^{-1} h^{-1}]. Note that the stoichiometric coefficients γ_{ij} are contained in the so-called stoichiometric matrix $\boldsymbol{S}$. Therefore, the matrix $\boldsymbol{S}$ itself can be used for the intracellular mass balance formulation:

$$\frac{dc}{dt} = \boldsymbol{S}r - \mu c, \tag{4}$$

where r is a flux vector, μ is the growth rate, and c is the concentration vector, which contains the concentrations for all components

(C_A, C_B, etc.). In most cases, the dilution term ($\mu\ c$) is small in comparison to the intracellular fluxes, and the equation can be simplified as follows:

$$\frac{dc}{dt} = \boldsymbol{S}\, r. \tag{5}$$

The steady state is a special case in which no temporal change of the intracellular concentrations is considered. This can be mathematically expressed as:

$$0 = \boldsymbol{S}\, r. \tag{6}$$

The equation above does not have a unique solution. The number of variables (reactions of the metabolic network) is usually much larger than the number of equations (metabolites), and measured reaction fluxes are normally scarce. Additional constraints can be applied to further reduce the number of allowable flux distributions [18]. Limits on the range of individual flux values can be used for this purpose; thermodynamic constraints [19] expressed as the directionality of a given reaction [16] can thus be used by setting one of the boundaries for that reaction to zero if the reaction is irreversible. In a similar way, maximum flux values can be estimated based on enzymatic capacity limitations [20], or for the case of exchange reactions, measured maximal uptake rates can be used (Sect. 3.3). Regulation of gene expression can also be considered in cases where the regulatory effects have a great influence on cellular behavior [21]. Usually, these constraints are not sufficient to reduce the solution space to a single solution. Therefore, linear programming methods are used to find a flux distribution that satisfies the problem:

$$\begin{aligned} &\max Z \\ &\text{subject to :} \\ &\quad \boldsymbol{S}r = 0 \\ &lb \leq r \leq ub, \end{aligned} \tag{7}$$

where Z is the objective function to be maximized (see Sect. 3.4.1), r is the flux vector, lb and ub are the lower and upper flux boundaries, respectively, and $\boldsymbol{S}$ is the stoichiometric matrix of the metabolic network. A reaction describing biomass generation [Eq. (1)] has been successfully used as an adequate objective function for predicting in vivo cellular behavior [22–24]. The above-explained approach is known as flux balance analysis (FBA) and is the most commonly used method for simulating the cellular phenotype.

2.2.2 Model for Bioreactor System

A generic mass balance equation for any component in a bioreactor can verbally be formulated as:

$$\begin{aligned}\text{accumulation of component} = {} & \text{mass added to the system} \\ & - \text{mass extracted from the system} \\ & + \text{mass converted in the system.}\end{aligned} \tag{8}$$

The term "mass converted in the system" refers to the catalytic activity of living cells. The equation is used to formulate mass balance equations for the volume of the reactor, for the biomass, and for the components in the liquid phase of the reactor. Table 2 gives an overview of the variables used. For a more complete description, refer to Kremling [25].

Volume of the Bioreactor

The dynamics of the reactor volume can be described by:

$$\frac{dm_R}{dt} = \sum q_{\text{in},\,j}\rho - q_{\text{out}}\rho. \tag{9}$$

If ρ is assumed to be constant and since $m_R = V_R\rho$:

$$\boxed{\frac{dV_R}{dt} = \sum q_{\text{in},\,j} - q_{\text{out}}.} \tag{10}$$

Table 2
Overview of the used variables and units for the reactor system

Name	Symbol	Units
Density	ρ	g l^{-1}
Reactor volume	V_R	l
Growth rate	μ	h^{-1}
Biomass yield on substrate i	Υ_{XS}	g g^{-1}
Volumetric feed j	$q_{\text{in},j}$	l h^{-1}
Volumetric reactor efflux	q_{out}	l h^{-1}
Mass of liquid in reactor	m_R	g
Biomass	m_B	g
Mass of component i	m_{Si}	g
Biomass concentration	c_B	g l^{-1}
Concentration of component i	c_{Si}	g l^{-1}
Molecular weight of component i	w_{Si}	g mol^{-1}
Exchange reaction for component i	r_{Si}^{e}	mol gDW^{-1} h^{-1}

Biomass

For modeling of the biomass, it is assumed that the feed contains no cells. Cell recirculation is also not considered. The mass balance for the biomass reads:

$$\frac{dm_B}{dt} = \mu m_B - q_{\text{out}} c_B. \tag{11}$$

The growth rate μ can be typically expressed as a function of the substrate uptake rate and the biomass yield: $\mu = \Upsilon_{X/S}\ r^{e}_{Si}\ w_{Si}$. For convenience, the biomass dynamics are now expressed in terms of biomass concentration. This is done by expressing the biomass in the reactor [g] as a function of the biomass concentration [g l^{-1}] and the reactor volume [l]:

$$\frac{dm_B}{dt} = \frac{d(V_R c_B)}{dt} = V_R \frac{dc_B}{dt} + \frac{dV_R}{dt} c_B. \tag{12}$$

Substituting Eqs. (12) and (10) into Eq. (11) and solving for biomass concentration lead to:

$$V_R \frac{dc_B}{dt} + \frac{dV_R}{dt} c_B = \mu m_B - q_{\text{out}} c_B$$

$$\boxed{\frac{dc_B}{dt} = \mu c_B - \frac{\sum q_{\text{in},j}}{V_R} c_B.} \tag{13}$$

Components in the Liquid Phase

The mass balance for substances (substrates/products) in the liquid phase is derived in a similar way as for the biomass. The mass balance for the component i is shown in Eq. (14). In this case, exchange reactions $r_{Si}{}^{e}$ between the cell and culture media have to be considered. A positive sign is used for products secreted by the cell, whereas a negative sign precedes $r_{Si}{}^{e}$ for substrates absorbed by the cell.

$$\frac{dm_{Si}}{dt} = q_{\text{in},j}\ c^{in}_{Si} - \ q_{\text{out}} c_{Si} \pm \ r^{e}_{Si} c_B\ V_R\ w_{Si}$$

$$\boxed{\frac{dc_{Si}}{dt} = \frac{q_{\text{in},j}}{V_R} c^{in}_{Si} - \frac{\sum q_{\text{in},j}}{V_R} c_{Si} \pm \ r^{e}_{Si}\ c_B\ \ w_{Si}}. \tag{14}$$

The mass balance equations derived for biomass, reactor volume, and components in the liquid phase can be used to describe the dynamics of a continuous ($q_{\text{in},j} \neq 0$; $q_{\text{out}} \neq 0$), a batch ($q_{\text{in},j} = q_{\text{out}} = 0$), or a fed-batch process ($q_{\text{in},j} \neq 0$, $q_{\text{out}} = 0$).

2.3 Experimental Data

With the development of high-throughput technologies, it is currently possible to produce large amounts of experimental data to characterize the proteome, genome, metabolome, and transcriptome of a microorganism under specific conditions. This allows a system-wide analysis of the cell response to genetic perturbations and operating conditions in the bioreactor, such as glucose and oxygen concentrations. Genome-scale reconstructions provide a suitable framework for the analysis and integration of these large datasets. To this end, many approaches have been developed over the last years. Hyduke [26] and Kim and Lun [27] provide a good overview of the possibilities of integrating omics data with genome-scale models. A recent multi-scale, genome-wide model of *E. coli* [28] represents an illustrative example of integrative modeling. The model incorporates the gene expression data of 4,189 genes in 2,198 conditions, transcriptional regulation, signal transduction, and metabolic pathways.

If the abovementioned high-throughput measurements are not readily available for the organism of interest, insights into the metabolism of wild-type and mutant strains can be gained using simple experiments. For instance, measurements of the time course of concentrations of extracellular metabolites can be used to determine cell-specific uptake and production rates [29]. The resulting rates can then be used as constraints for the corresponding exchange reactions used to reduce the solution space of the metabolic model describing the metabolism of the cell (see Sects. 2.2.1 and 3.3).

2.4 Software

Table 3 summarizes some commonly used software packages that support the calculations necessary for Metabolic Engineering. Some tools, like YANA, are stand-alone and need no extra software for their operation. Some others, like the widely used COBRA Toolbox, are packages that require previous installation of a specific platform (Matlab or Python) and a solver. Python + COBRApy + Glpk represent high-quality, free, open-source options and are recommended if a Matlab license is not available. Gurobi offers a free academic license and is therefore a good option when performing quadratic or quadratically constrained programming.

2.5 Next-Generation Models for Metabolic Engineering

Metabolic processes taking place in the cell are strictly coordinated by highly interconnected, complex, and sometimes intricate networks. The activity level of a specific enzyme in the cell can be regulated at the transcription/translation level as well as by using posttranslational modifications, which in turn are coordinated by signaling networks. Thus, observable cellular behavior results from a complex interplay of multiple cellular networks. First attempts to integrate metabolic reconstructions into additional networks have already been made by many research groups [28, 35–37].

Table 3
Commonly used software for calculations in Metabolic Engineering

		Description	Reference/URL
Platform	*Matlab*	High-level language for numerical computation, visualization, and application development	www.mathworks.com
	Python	High-level, multi-paradigm programming language. It is a free and open-source software and has a community-based development model	www.python.org
	Mathematica	Computational software program used in many scientific, engineering, mathematical, and computing fields, based on symbolic mathematics	www.wolfram.com/mathematica
Toolbox	*COBRA for Matlab*	Matlab package for implementing COBRA (constraint-based reconstruction and analysis) methods to simulate, analyze, and predict a variety of metabolic phenotypes using genome-scale models	[30]
	COBRA for Python (COBRApy)	Python package that provides support for basic COBRA methods. COBRApy includes parallel processing support for computationally intensive processes	[31]
	CellNetAnalyzer	Matlab toolbox that provides a graphical user interface and various computational methods and algorithms for exploring structural and functional properties of metabolic, signaling, and regulatory networks	[32]
	Pathway Pioneer	Web-based biological engineering tool that allows dynamic interaction with biological models. The underlying data is flux balance analysis (FBA) computed using COBRApy	www.pathwaypioneer.org
	SNA: stoichiometric network analysis	Interactive, high-performance toolbox for analyzing steady-state behavior of metabolic networks. The toolbox is mainly implemented in Mathematica	[33]
	YANA	Platform-independent, dedicated toolbox for metabolic networks with graphical user interface to calculate, edit visualize, centralize, and compare elementary flux modes	[34]
Solver	*Glpk*	The GNU Linear Programming Kit (Glpk) is intended for solving large-scale linear programming (LP), mixed integer programming (MILP), and other related problems	www.gnu.org/software/glpk
	Gurobi	Commercial solver for optimization problems. Free academic license available. Supports LP, quadratic and quadratically constrained programming (QP and QCP), and MILP	www.gurobi.com
	Lindo	Commercial optimization modeling software	www.lindo.com
	Mosek	Tool for solving mathematical optimization problems: LP, QP, conic problems, MILP	www.mosek.com

A current example is the development of the first model, which aims to integrate metabolism and gene expression (ME-Models) for *E. coli* [35, 36]. ME-Models extend the prediction capabilities of the traditional metabolic models (M-models), allowing, for instance, the assessment of the metabolic burden observed in cells expressing large engineered pathways. Thus, with ME-Models, engineering strategies to overcome the metabolic burden can be better explored. With the addition of further details and the refining of the ME-Models [37], the dimension of the stoichiometric matrix grows to a computationally challenging magnitude. The great scope of the ME-Models encompasses not only their tractability but also their analysis of the simulation results. As an alternative to these detailed models, a mechanistic ODE-model (compartment model) that describes transcription and translation [38, 39] of gene pools can be coupled with a metabolic model. The application of such a compartment model that describes the relationship between growth rate and the content of RNA, DNA, and bulk protein, and additionally accounts for the amount of free and bounded ribosomes, improves the prediction capabilities of the extended model while keeping it tractable.

3 Methods

Here, we propose a five-step Metabolic Engineering strategy to achieve the optimal production of a target molecule in a selected host microorganism. Figure 3 summarizes the main phases of the strategy and shows the associated chapters, in which each step is explained in detail.

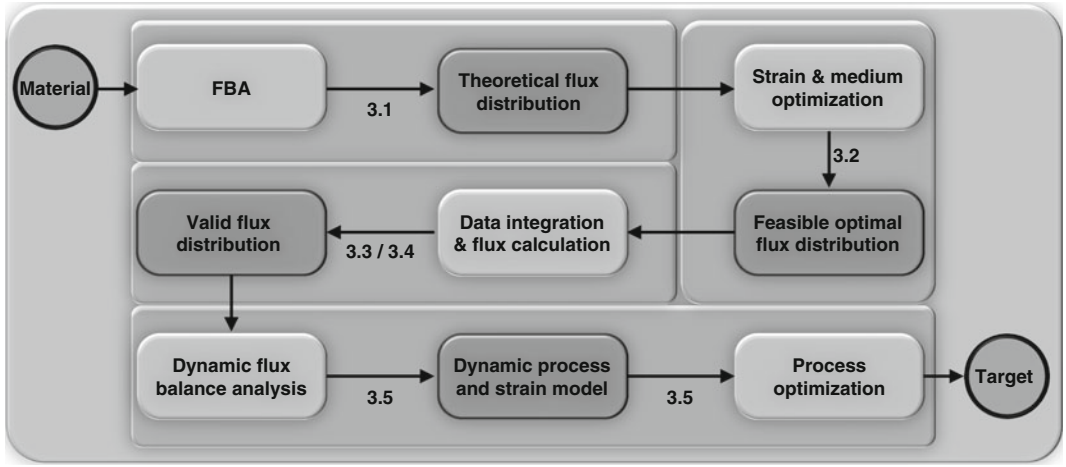


Fig. 3 Five-step Metabolic Engineering strategy. *Light gray symbols* represent methods, while the *dark gray symbols* represent the result of these methods. Material is an input in different stages but is shown only for the first method. The sections of this book chapter are represented by the *numbers*

In the first step, a theoretical characterization of the capabilities of the strain is performed. Moreover, optimal pathway configuration and medium composition are estimated for the production of the target. For this purpose, an adequate metabolic model is analyzed using flux balance analysis (FBA) and its extensions. In the second step, in silico strain optimization algorithms are used to predict gene/reaction deletions that redirect the carbon flow toward the production pathways. In the third and fourth steps, experimental data is analyzed and integrated. The comparison of the estimated intracellular flux pattern of the wild-type and mutant strains can be used to evaluate the effect of genetic manipulations on the improvement of product yield. In the last step, the performance of the engineered strains in a bioreactor is assessed, and by selecting adequate process conditions, improvements of productivity and final titer are achieved. Dynamic flux balance analysis plays a central role in this last step.

3.1 Theoretical Product Yields and Pathways

Even before experimental data for the strain to be engineered is available (Sect. 2.3), a pure theoretical characterization of the metabolic system capabilities can be performed using a suitable metabolic reconstruction. The methods discussed in this section include:

- *Theoretical product yields:* can be used as an indicator for the performance potential of the wild-type/mutant strains under different conditions
- *Optimal pathway configuration:* facilitates decisions about which pathway or pathway combinations have to be used to optimally produce the target molecule
- *Optimal medium composition:* guides the selection of the real medium composition by showing which substrates have a positive impact on product yield

3.1.1 Calculation of the Theoretical Product Yields

A metabolic reconstruction, specific for the host strain used, is necessary to calculate the maximal theoretical product yield supported by the host microorganism. The theoretical product yield is a function of the thermodynamic, stoichiometric, and physiological constraints considered when performing the calculations. The procedure for calculating the maximal theoretical yield is explained for the production of succinate in *E. coli* as a case study.

1. Choose a metabolic reconstruction of *E. coli*. See Table 1.
2. Set the production of succinate as an objective function. Here one can choose between selecting an existing reaction and adding a new one to the model. In case of the *E. coli* core model, the reaction *SUCCt3* (succinate transport out via proton antiport) is a good candidate for the objective function. If one decides to add a new reaction, it should be of the form "succ[c] →."

3. Define the medium composition. This is done by modifying the upper and lower limits of the exchange reactions. For this specific example, we will assume that glucose is the sole carbon source.
4. Add additional constraints to the model (gene deletions, growth rate, GAM value, etc.).
5. Assume an arbitrary uptake rate for glucose (if no experimental measurements are available) and solve the linear programming problem using an adequate solver (*see* **Note** 1).
6. The resulting flux distribution should now be scaled to the input flux of glucose in order to get the value of the maximal theoretical yield (*see* **Note** 2).

The COBRA Toolbox provides a set of functions that facilitate the execution of all these steps with only a few code lines (Table 4). For a detailed explanation of these functions, refer to the COBRA Protocol [30].

The effect of imposing different constraints on the maximal theoretical yield is illustrated in Fig. 4. Aerobic and anaerobic cultivations are considered with glucose as the sole carbon source. Three situations are analyzed: two cases in which growth is not considered and one case in which the growth rate has an arbitrary value of 0.35 h^{-1}. The yield values reported in Fig. 4 represent the limits for the metabolic system under these conditions. No higher yields are possible as long as the metabolic network is not modified (addition or stoichiometry modification of reactions). It can be seen that growth has a negative effect on the maximal theoretical yield.

Table 4
Theoretical maximal yield calculations using the core model of the *E. coli* metabolism and functions of the COBRA Toolbox

Matlab code	Explanation
`model=readCbModel('ecoli_core_model.xml');`	Load the *E. coli* core model
`model=changeObjective(model,'SUCCt3');`	Set the objective function
`model=changeRxnBounds(model,'EX_glc(e)',-1,'l');`	Assume an uptake rate for glucose
`model=changeRxnBounds(model,'EX_o2(e)',0,'b');`	Define the medium composition, e.g., oxygen
`model=changeRxnBounds(model,'SUCCt2_2',0,'b');`	Constrain the solution to avoid cycles
`model=changeRxnBounds(model,'ATPM',0,'b');`	Optional: assume no maintenance ATP requirement
`solution=optimizeCbModel(model,'max');`	Optimize the LP problem

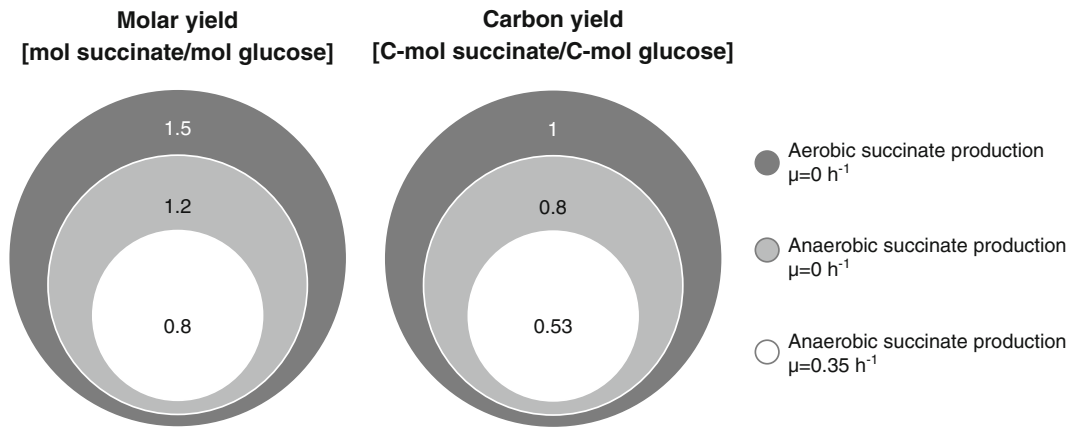


Fig. 4 Effect of constraints on theoretical yields: aerobic and anaerobic cultivations are considered with glucose as sole carbon source for the production of succinate in *E. coli*. Three different situations are analyzed: two cases in which growth is not considered ($\mu = 0$) and one case with a growth rate of 0.35 h^{-1}. The maximal possible yields are calculated for each situation, and the results are presented as molar (*left*) and carbon (*right*) yields. A carbon yield of 1 (equivalent to a molar yield of 1.5) indicates that the metabolic system is capable of converting all supplied carbon atoms into product

This is a logical consequence if the biomass is considered as an additional product that has to be synthetized by the metabolic system. The more biomass is produced, the less carbon will be available for the production of the target biomolecule. Additionally, Fig. 4 shows that for the succinate production in *E. coli*, a maximal theoretical carbon yield of one can only be reached under aerobic conditions. This is further analyzed in Sect. 3.1.3.

3.1.2 Determining an Optimal Pathway Configuration Using Stoichiometric Analysis

Many bio-products can be produced using different biochemical routes. These routes can occur naturally either in the host strain itself (native pathways) or in other organisms (heterologous pathway), or they can be synthetically generated. Metabolic engineers are thus often confronted with the task of selecting the best pathway configuration to be engineered in the host cell. Pathway configuration refers here to the situation of using pathway A, pathway B, or a combination of both for the biosynthesis of a target product R (Fig. 5a). This choice should be made considering many aspects, e.g., energy, cofactor, and reduction equivalent consumption. Curated metabolic models can be used to guide the selection process of the pathway configuration with the best performance index: molar and carbon yield are commonly used performance indices when comparing pathways (*see* **Note** 3). The procedure of finding a pathway configuration that reaches the maximal performance index is explained using a hypothetical case study, in which pathway A and pathway B lead to the formation of the product R. In the hypothetical case study, pathway A is a native pathway, while pathway B is a heterologous one.

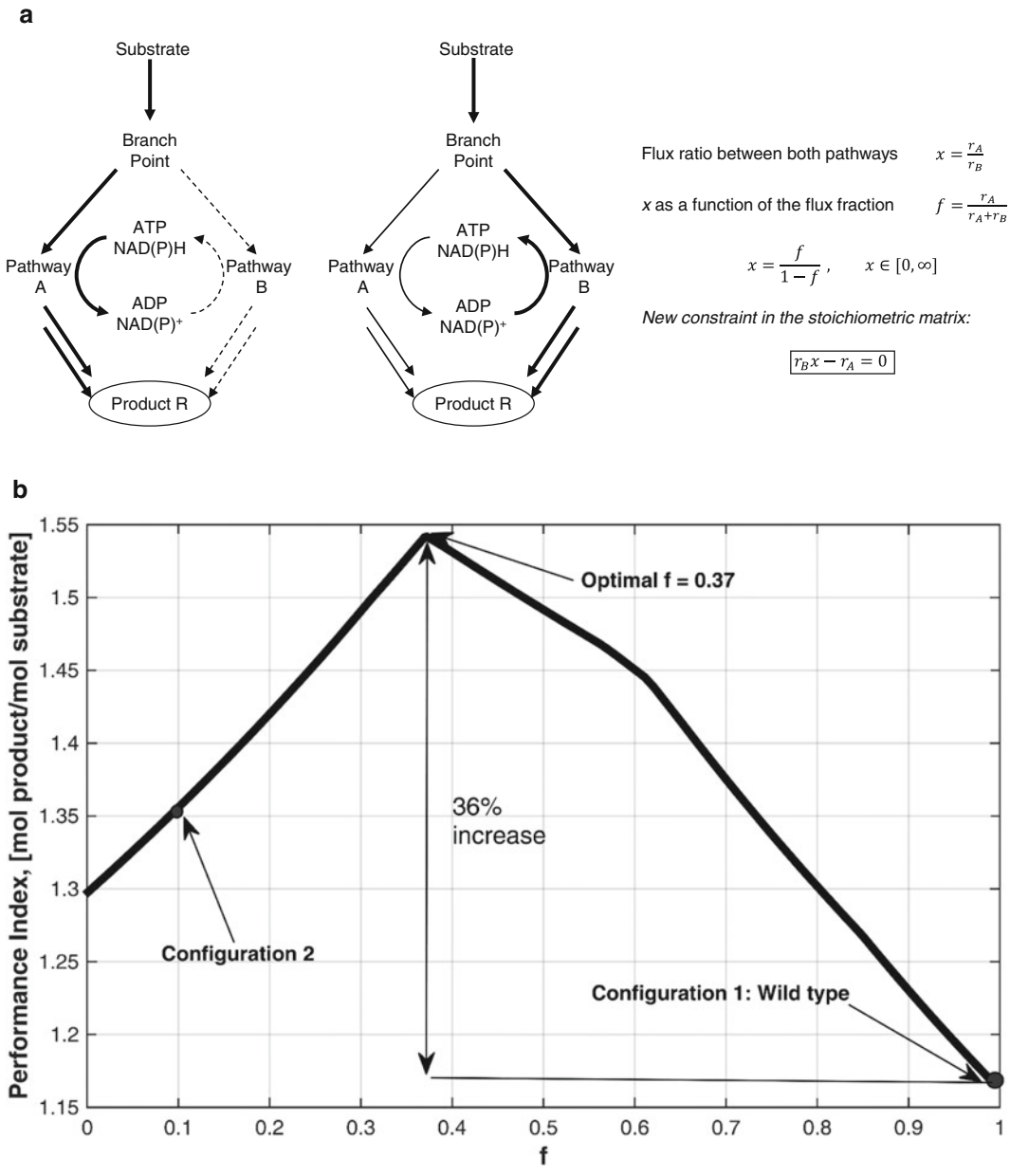


Fig. 5 (**a**) Pathway configuration for wild-type and engineered but suboptimal strain: two different situations are shown. The *left panel* shows the wild type, in which the substrate flows only into the native pathway A ($f = 1$). The native pathway A consumes one molecule of ATP and NAD(P)H. In the *central panel*, pathway B is added to the wild-type strain. The heterologous pathway B produces one molecule of ATP and NAD(P)H. In this engineered strain, 10% of the carbon flows into the native pathway A and the rest into pathway B ($f = 0.1$). This pathway configuration is not optimal. Equations describing the carbon distribution are shown in the *right panel*. r_A and r_B represent the rates through the first reaction of pathways A and B, respectively. A *f*-value of one means that all of the carbon flows into pathway A. (**b**) Simulation of different pathway configurations. The maximal performance index is reached when 37% of the substrate flows into the native pathway A ($f = 0.37$). The synergy observed arises from the dynamics of ATP and NADPH between the two pathways

1. Identify the different pathways that lead to the target bioproduct formation.
2. Incorporate new biochemical routes, if necessary.
3. Identify the common branch point of the pathways. See Fig. 5a.
4. Use the flux ratio between pathway A and B to specify the flux through each pathway. Use the variables x and f as shown in Fig. 5a for this propose.
5. Add the marked equation in Fig. 5a as a new constraint in the stoichiometric matrix. The coefficients are x and -1 for the first reaction of pathways A and B, respectively.
6. Calculate the maximal theoretical performance index for different pathway configurations, i.e., different values of flux ratio x or flux fraction f. Note that the variable f (flux fraction) can only take values between 0 and 1, while the flux ratio ranges from 0 to infinity.
7. Select the optimal pathway configuration from the simulation results. See Fig. 5b.

The effect of the pathway configuration on the selected performance index, in this case molar yield, is illustrated in Fig. 5b. In this hypothetical case study, the native pathway A has a lower performance index than the heterologous pathway B. However, the system only becomes optimal when both pathways are expressed in a fraction of $f = 0.37$. Shen and Liao [40] experimentally proved the validity of the approach described above when engineering *E. coli* for the production of 1-propanol. They observed an improvement of the 1-propanol yield of 30–50% when expressing both the heterologous citramalate pathway and the native threonine pathway for the production of the 1-propanol intermediate 2-ketobutyrate, compared to the yield when using only one pathway. The synergy observed was in good agreement with the predictions made with the approach explained above.

3.1.3 Estimation of the Culture Medium Composition

Which are the optimal substrates for the production of a desired target molecule? Should the production be performed under aerobic or anaerobic conditions? Can the totality of the assimilated carbon be transformed into product by the metabolic system in its actual configuration? Finding the answers to these and similar questions can be challenging and requires a great experimental effort. However, when dealing with these issues, a sensitivity analysis of the metabolic model of the strain being engineered can be helpful. The general procedure is illustrated again, using the example of succinate production in *E. coli*.

1. Select an adequate metabolic reconstruction of the analyzed strain. See Sect. 2.1.

2. Define a base model. This model will represent the starting conditions for the sensitivity analysis. In the concrete case of the succinate production in *E. coli*, the starting conditions correspond to no carbon source and anaerobic conditions.
3. Define a biologically meaningful range for each analyzed exchange reaction. For instance, the glucose uptake rate was assumed to have a maximal value of 18 mmol gDW^{-1} h^{-1}.
4. Vary the lower limit of an arbitrary exchange reaction, inside of the predefined range, in order to permit the system to absorb the corresponding compound.
5. Calculate the maximal value of the desired performance index. See Sect. 3.1.1.
6. Repeat steps 4 to 5 until all exchange reactions of interest are analyzed.
7. Sort the results in a table and make decisions about what compound should be added to the culture medium (base model) in order to improve the performance index.
8. If the desired value for the performance index is not reached after modifying the base model, repeat steps 1–7 until the desired performance is accomplished.

The production of succinate in *E. coli* is an extensively studied process. Therefore it is a good case study to show the utility of the approach explained above in guiding the selection of medium composition. Encouraging steps toward an engineered *E. coli* strain with high yield, productivity, and titer have been made. Most of the work reported to enhance the succinate production has been performed under anaerobic conditions [41–43]. Interestingly, a simple sensitivity analysis shows (Table 5b) that the maximal theoretical carbon yield can only be reached under aerobic conditions. *E. coli* mutants, which produce succinate under aerobic conditions, and a theoretically designed high-performance aerobic strain have been reported [44, 45]. The sensitivity analysis can also guide the development of an anaerobic cultivation process. It shows that the addition of carbon dioxide to the system has a positive effect on the maximal yield. It is therefore logical to use a carbon dioxide atmosphere in anaerobic cultivations. This fact has been identified and used by many research groups [46–48].

3.2 In Silico Strain Optimization

Genome-scale reconstructions of metabolism have been used for over a decade now to predict genetic modifications that improve the product yield and production performance of the engineered production strains. Some manipulation strategies that can be explored in silico are listed in Table 6. A more extensive overview of these methods can be found in [5]. The first algorithms for in silico strain design permit us to predict the effect of reaction

Table 5
Sensitivity analysis of succinate production in *E. coli*. The core model was used for the calculations. (a) First round of the sensitivity analysis. The maximal theoretical carbon yield [C-mole succinate/C-mole glucose] for each carbon source is shown. Malate and fumarate exhibit the highest performance, followed by fructose and glucose. (b) Glucose is selected as the carbon source and a second round of sensitivity analysis is performed. The addition of oxygen permits the model to reach the maximal possible carbon yield. Under these conditions it is theoretically possible to use all carbon atoms of glucose for the synthesis of succinate (c). After addition of oxygen, the system has reached its optimum, and further modifications have no effect on the yield

(a) Base model: no carbon source, anaerobic, $\mu = 0$		(b) Base model: glucose, anaerobic, $\mu = 0$		(c) Base model: glucose, aerobic, $\mu = 0$	
EX_mal-L (e)	0.85714	EX_o2 (e)	1	EX_ac (e)	1
EX_fum (e)	0.85714	EX_fum (e)	0.99966	EX_acald (e)	1
EX_fru (e)	0.8	EX_mal-L (e)	0.99966	EX_akg (e)	1
EX_glc (e)	**0.8**	EX_co2 (e)	0.98778	EX_co2 (e)	1
EX_pyr (e)	0.44444	EX_akg (e)	0.97978	EX_etoh (e)	1
EX_acald (e)	0.4	EX_pyr (e)	0.87516	EX_for (e)	1
EX_akg (e)	0.4	EX_gln-L (e)	0.85674	EX_fru (e)	1
EX_lac-D (e)	0.23529	EX_glu-L (e)	0.85674	EX_fum (e)	1
EX_ac (e)	0	EX_ac (e)	0.8	**EX_glc (e)**	**1**
EX_co2 (e)	0	EX_fru (e)	0.8	EX_gln-L (e)	1
EX_etoh (e)	0	**EX_glc (e)**	**0.8**	EX_glu-L (e)	1
EX_for (e)	0	EX_lac-D (e)	0.8	EX_h2o (e)	1
EX_gln-L (e)	0	EX_acald (e)	0.8	EX_h (e)	1
EX_glu-L (e)	0	EX_etoh (e)	0.8	EX_lac-D (e)	1
EX_h2o (e)	0	EX_for (e)	0.8	EX_mal-L (e)	1
EX_h (e)	0	EX_h2o (e)	0.8	EX_nh4 (e)	1
EX_nh4 (e)	0	EX_h (e)	0.8	EX_o2 (e)	1
EX_o2 (e)	0	EX_nh4 (e)	0.8	EX_pi (e)	1
EX_pi (e)	0	EX_pi (e)	0.8	EX_pyr (e)	1
EX_succ (e)	0	EX_succ (e)	0.8	EX_succ (e)	1

EX_mal-L(e) exchange reaction for L-malate, *ac* acetate, *acald* acetaldehyde, *akg* 2-oxoglutarate, *etoh* ethanol, *for* formate, *fru* D-fructose, *fum* fumarate, *glc* D-glucose, *gln* L-glutamine, *glu* L-glutamate, *h* H^+, *lac-D* D-lactate, *mal-L* L-malate, *nh4* ammonia, *pi* phosphate, *pyr* pyruvate, *succ* succinate

knockouts, that is, they consider the effect that reaction deletions have on the metabolic network and product yield. Predicting the up- and downregulation of reactions represents an extension of these first algorithms. Additionally, if gene-protein reaction

Table 6
Common strain optimization algorithms

Strategy	Algorithm	Approach	Reference
Reaction knockout	OptKnock	Bilevel optimization, MILP	[49]
	RobustKnock	Bilevel max-min optimization, MILP	[50]
Gene knockout	OptGene	Genetic algorithm	[51]
	GDLS	MILP	[52]
Reaction upregulation/ downregulation	EMILiO	Bilevel optimization, iterative linear program (ILP), and MILP	[44]
	OptForce	LP	[53]

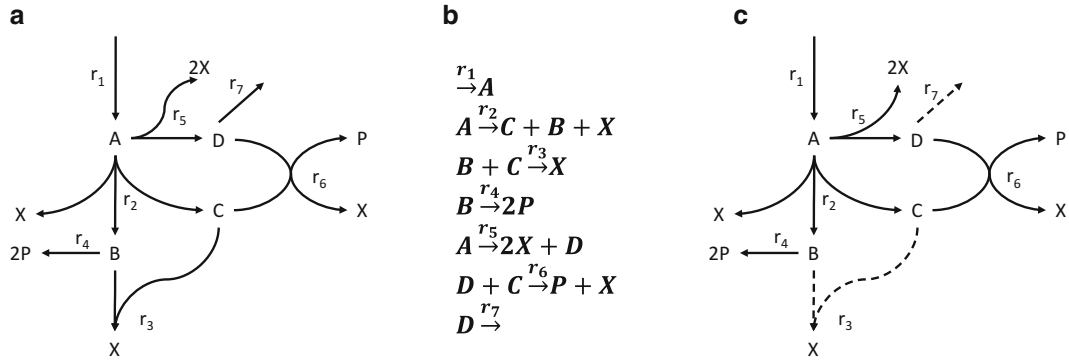


Fig. 6 (**a**) Hypothetical metabolic network for the production of P and X. (**b**) The network consists of seven reactions. A, B, C, and D are intermediates. (**c**) Two reaction knockouts – r_3 and r_7 – are necessary to maximize the production of biomass and at the same time product P [25]

(GPR) mappings are available for the metabolic system being analyzed, algorithms that only predict gene knockouts can be used and should be preferred, as exactly these modifications will later be experimentally implemented in the real biological systems.

The general procedure used to perform in silico strain optimization is explained using the metabolic network shown in Fig. 6, as described by Kremling [25]. X and P represent the biomass and the target product, respectively. P, C, and B are intermediates. In this example, the synthesis of P is coupled to growth (X).

1. Set up the metabolic model or choose an existing genome-scale metabolic reconstruction.
2. Identify the reactions that contribute to the production of the target molecules. In this example the reactions r_4 and r_6 synthetize the product P, and the reactions r_2, r_3, r_5, and r_6 contribute to the formation of biomass.

3. Define the objective functions: $Z_1 = f(r_4,r_6)$ and $Z_2 = f(r_2,r_3, r_5,r_6)$ for product and biomass, respectively.
4. Select the set of reactions that can be deleted from the network.
5. Determine the number, n, of reactions to be knocked out. The computing time required to find a solution depends on the algorithm used and can increase exponentially or linearly with the total number of knockouts in the mutant strain.
6. Select a strain optimization algorithm (Table 6) and perform the simulation. It is strongly recommended to use more than one algorithm to perform the in silico strain optimization. Since each algorithm examines the solution space in a different way (local/global search, one path/multiple path), finding different solutions is to be expected. OptKnock is a good starting point and is already implemented in the COBRA Toolbox.
7. Analyze the predicted gene deletions in respect to biological consistence and select the best option to be experimentally implemented. Note that due to inherent inaccuracies in the metabolic model, not all predicted mutants are biologically feasible. Further genetic modifications might be necessary in order to obtain the optimal flux distribution that maximizes the product yield. A good example for this situation is the design of a high-performance aerobic *E. coli* strain for the succinate production, in which additional genetic modifications are necessary to obtain the optimal flux distribution predicted by EMILiO [44].

In the case of the network shown in Fig. 6, only two reaction knockouts are sufficient to maximize the reaction flux through the product and biomass. The network was optimized using the OptKnock algorithm. Reaction two to reaction seven conform the set of reactions that can be deleted. The calculation time required was 0.005 s.

3.3 Analysis of Experimental Data

For analyzing and optimizing a host strain, substrate uptake rates and product excretion rates must be determined to calculate the complete flux distribution. The rates measured can be used to identify bottlenecks as well as to confirm engineering success. Basically, the more metabolic data are available, the more precise and better is the evaluation of the fluxes.

Biomass and substrate concentrations especially are easily measurable during an experiment, and commercial kits for accessing them are often available. In many laboratories, measurement tools like HPLC to quantify the cellular output in the form of metabolites have already been established. In an open system, such as the standard shaking flask, it is not possible to close the mass balance because of the impossibility of determining all carbon fluxes in

the system (e.g., CO_2 that is produced by the cells). Therefore, a bioreactor system and an exhaust gas analyzing system are recommended.

After pre-culturing, the strain of interest should be inoculated in a defined minimal medium with a carbon source of interest. According to the fermentation strategy, either a feeding strategy or a batch cultivation with a specific initial concentration of the carbon source can be applied. Substrate feeding has to be included in the mathematical analysis as described in 2.2.2. The specific uptake and formation rates are determined depending on the corresponding time course data for the metabolite concentrations in the bioreactor system. To determine all relevant rates, data for all metabolites and for cell dry mass have to be taken from the same time frame and growth phase as shown in the gray-shaded area in Fig. 7a.

As the rates should be normalized to the cell dry mass (DW), as a first step it is useful to correlate the measured optical density (OD) with the biomass (Fig. 7b). If there is no DW available, the OD can

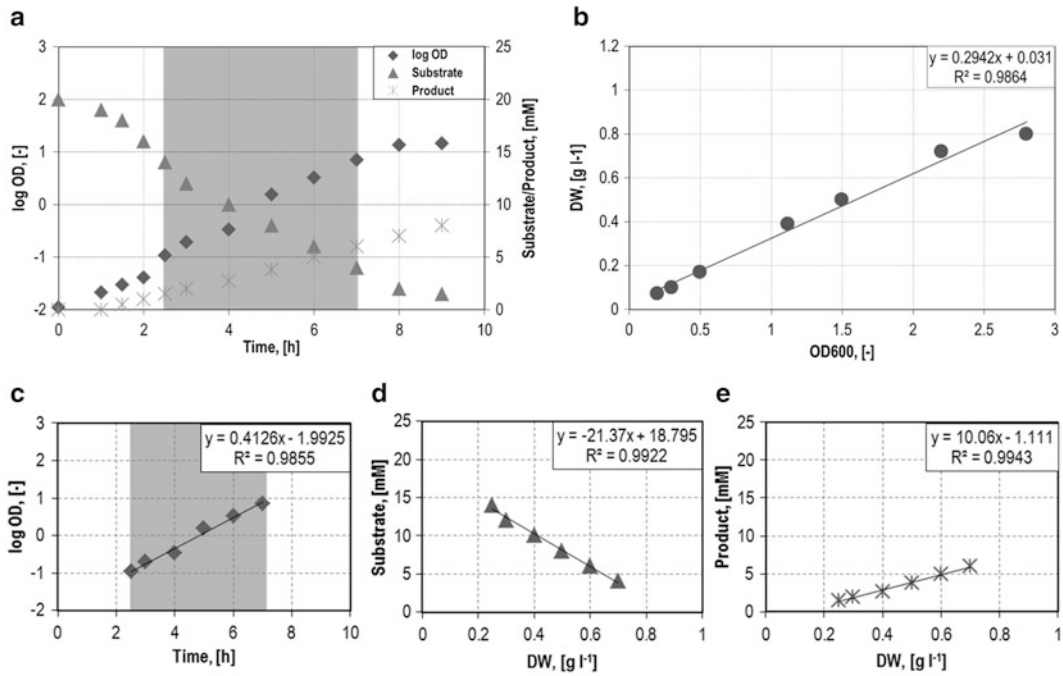


Fig. 7 Fictive data for strain performance during a production process (**a**). The measured optical density increases with rising biomass (**b**). Graph **c** shows the logarithmic optical density in the exponential phase against time. Substrate concentration (**d**) and product concentration (**e**) are plotted with respect to the cell dry weight. To evaluate the respective rates, concentrations are related to biomass formation. Therefore, equations of the contemplated gradients are necessary. The rate is equal to the slope of the straight line multiplied by the growth rate. Corresponding rates have to be analyzed in the same time frame and growth phase (*gray shaded*). A high coefficient of correlation R^2 is required to obtain conclusive results

also be used as a proxy for biomass. In order to correlate substrate uptake and product formation rates, it is important to always use the same parameters (wavelength, growth stage, medium, etc.). At least three biological replicates should be measured to minimize the standard deviation of the measurement.

The biomass formation, also known as specific growth rate μ, is one particular case of formation rate. It can be determined directly from the slope of the measured logarithmic OD curve. Due to the (linear) correlation between OD and biomass concentration (Fig. 7b), the slope of the data points in the chosen interval is equal to the growth rate μ [Fig. 7c, Eq. (15)].

Metabolic rates have to be determined from the extracellular time course data of substance depletion or accumulation $\frac{dc_{si}}{dt}$ that has to be related to the currently measured biomass concentration c_B [Eq. (16)]. Uptake and formation rates r for a substance S not only correspond to the respective time point of the measurement but also to the already generated biomass c_{Bt} [29]. The temporal alteration of biomass [Eq. (15)] has to be linked to the uptake or formation rate [Eq. (16)] to analyze those rates:

$$\mu = \frac{1}{c_{Bt}} * \frac{dc_B}{dt} \quad \text{resp.} \quad c_{Bt} = \frac{1}{\mu} * \frac{dc_B}{dt} \tag{15}$$

$$r = \frac{1}{c_B} * \frac{dc_{si}}{dt} = \frac{\mu * dt}{dc_B} * \frac{dc_{si}}{dt} = \mu * \frac{dc_{si}}{dc_B}. \tag{16}$$

In order to calculate the substrate uptake rate as well as the product formation rate, the according concentration has to be plotted against the corresponding biomass (Fig. 7d and e). The slope $\frac{dc_{si}}{dc_B}$ multiplied by the growth rate μ gives the respective rate [Eq. (16)] (*see* **Note** 4).

3.4 Estimation of the In Vivo Flux Distribution

First of all, measured rates can be fed into the already established metabolic model to restrict the solution space (Fig. 8). Additionally, to provide realistic flux estimations, an objective function and an adjusted value for GAM (see below) are included in the calculations.

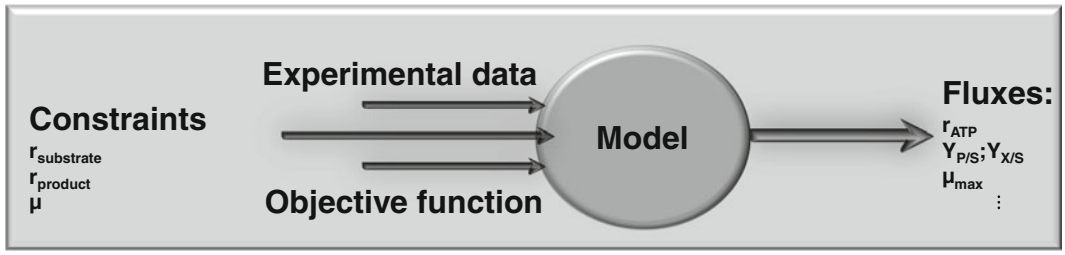


Fig. 8 In a comprehensive model, there are more unknown variables than equations. For defining a range of solutions, some constrains have to be given [11]. Additionally, a suitable objective function concerning the model output is required to compute an optimal network state and a resulting flux distribution [14]

3.4.1 Selecting an Adequate Objective Function

The most common assumption is that microbial cells maximize their growth [54]. For this reason biomass production is a frequently used objective function. However, depending on the growth phase, different objective functions as summarized in [54] are possible. So, in some cases it will be advantageous to combine a number of objective functions to restrict the solution space. If there is no growth, the energetic efficiency could be optimized instead of biomass yield. In this case the objective function would be the ATP yield. Another approach is to minimize the substrate consumption or the required number of reaction steps.

3.4.2 Calculation of Growth-Associated Maintenance (GAM)

Energy has to be considered for growth prediction. Cells have a specific energy requirement for maintenance metabolism [10]. This rate is defined as nongrowth-associated maintenance energy (NGAM) [14]. The yield of substrate uptake necessary for the resulting growth is defined as growth-associated maintenance metabolism (GAM) [14]. The available energy is specified with respect to the ATP concentration in order to meet different substrate compositions. The following steps need to be completed to calculate specific GAM values:

1. Determine growth rate and substrate uptake rates of the analyzed strain from different experimental setups (here named as setups 1–5) with various growth rates (e.g., adjusted by dilution rate in a continuous culture using a chemostat [11] (Fig. 9a)).
2. Compute the slope of the linear growth rate/substrate uptake rate correlation between $\mu = 0$ and μ_{max} (Fig. 9b).
3. Calculate, with the help of the already available stoichiometric model, different theoretical yields $Y_{X/S}$ under different theoretical GAMs (here named GAM1–GAM3) (Fig. 9c) as described above (Sect. 3.1.1). The strain-specific NGAM value is assumed to be the same.
4. After correlation of the used theoretical GAMs (here again GAM1–GAM3) with calculated yield $Y_{X/S}$ (Fig. 9d), the theoretical GAM out of the interpolated experimental data (Fig. 9b) can be read out as shown in Fig. 9e.
5. Depending on the model, wild-type *E. coli* strains have a GAM around 60 $mmol_{ATP}\ gDW^{-1}\ h^{-1}$ [16]. Mutants with an altered network structure will show a different behavior, and growth yields can be compared (Fig. 9f).
6. The existing model can be adapted to be strain specific with the experimentally determined GAM. For the modeled strain, it is possible to predict growth rate and flux distributions for a given substrate uptake as well as vice versa.

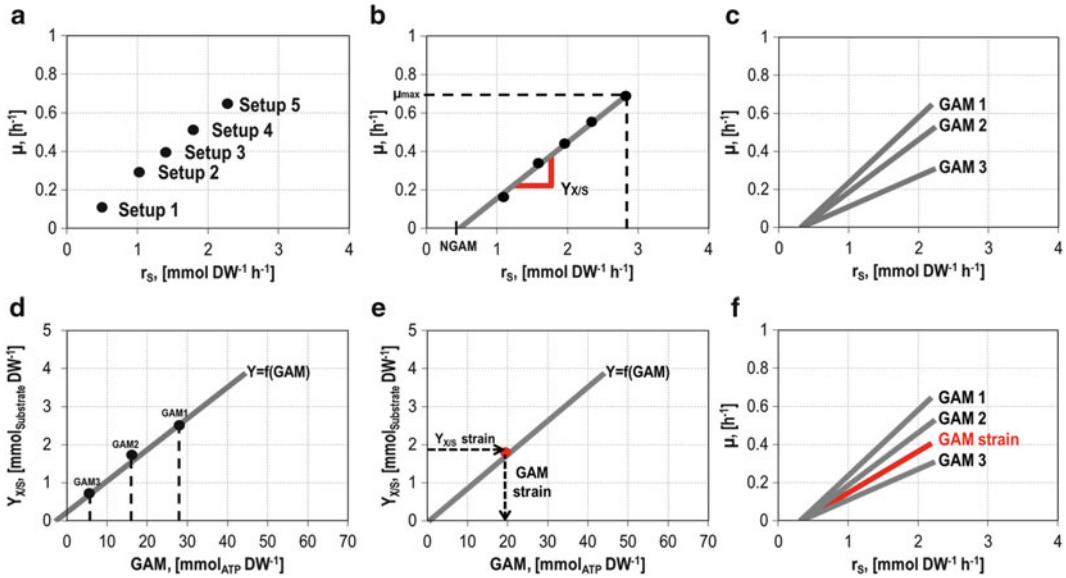


Fig. 9 The strain of interest can be analyzed by considering different growth rates and substrate uptake yields under various setups (**a**, **b**). Theoretical strain behavior with fixed GAM values has to be computed (**c**) to create the $Y_{X/S}$ to GAM ratio (**d**). From this correlation the GAM of the strain can be determined (**e**). The yield of the substrate with respect to the growth rate can be compared to the performance of other strains (**f**), and growth rates can be predicted

The value for NGAM can be calculated as described in Sect. 3.1.1 by setting ATP as an objective function and measured $q_{S\ NGAM}$ as input (Fig. 9b).

3.4.3 Reconstructing In Vivo Flux Distributions

The strain performance has to be recorded under various conditions and experimental setups with different growth rates to determine flux distributions. Flux balance analysis enables the calculation of the flux through the metabolic network of the cell (see above Sect. 3.1.1). For many organisms, these networks are already available online. It is possible to reconstruct the flux distribution in a microbial cell based on metabolic reconstructions in the systems biology markup language (SMBL) format available online with the help of the COBRA Toolbox [30]. This toolbox allows the visualization of the actual fluxes and offers us the opportunity to compare fluxes in mutant strains with wild-type flux distributions (*see* **Note** 5).

As shown in Fig. 10a, many rates in the *E. coli* core model are not available (thin arrows). It is possible to calculate the carbon flux in silico by setting the measured uptake and excretion rates as additional constraints. Figure 10b shows a hypothetical flux distribution through the network. Setting the hypothetical flux D–E to zero (e.g., by a mutation) results in a measurable shift in the production rate of metabolite I. If the predicted flux distributions are not congruent to the actual measured rates, it is an indication that regulatory interactions or further pathways are missing in the model.

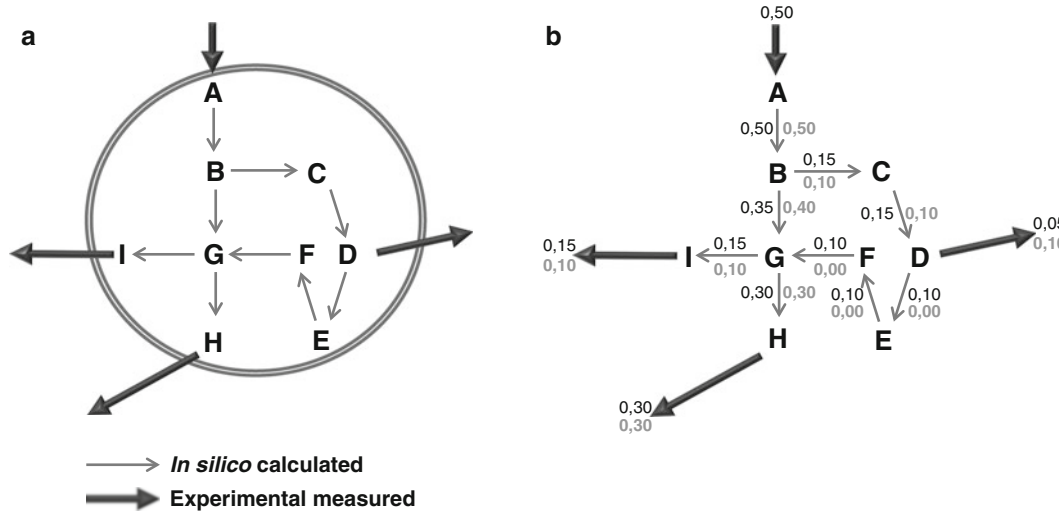


Fig. 10 Metabolic data can only partially be determined via experimental setups. As shown in a schematic cell with metabolites A–I, experimental measured rates (*thick arrows*) have to be supplemented by in silico calculated rates (*thin arrows*) for wild type and mutants to determine the complete flux distribution (**a**). *Dark numbers* in (**b**) represent the actual flux rates. A modification in the network, for example, the flux **D** → **E** is set to zero, results in an altered flux distribution as shown in (**b**). New rates are depicted in *light gray*

The presentation of metabolic networks is realized via graphical visualization of subnetworks. The output generated by Matlab enables a comparison of flux rates as shown in Fig. 10b. Maps from different metabolic pathways are available in BiGG knowledge base [55]. To create the image it is necessary to load the chosen map of interest into Matlab and draw it as a Matlab figure. This can be realized with only a few commands [30].

The stoichiometric matrix has to be modified to adapt the reconstruction to mutant strains proposed via in silico strain optimization. To get an idea of the most probable solution, the solution space of the mutant could be analyzed using the MOMA method, which is the minimization of metabolic adjustment [56]. This theory is based on the assumption that the "fitness" of the wild type has evolved over millions of years and represents the optimal metabolic state. This kind of pressure is not present for genetically modified organisms, which means that they probably do not possess the optimal growth configuration [56]. Caused by this, the most realistic flux distribution is the one derived from the solution space which contains the minimal distance to the wild-type flux distribution.

3.5 Assessing and Improving the Performance of Engineered Strains in a Bioreactor

The goal of the algorithms for strain optimization discussed in Sect. 3.2 is the redesign of the host metabolic network to maximize the yield of a target molecule while simultaneously supporting growth. This approach does not explicitly take into account the subsequent utilization of the engineered strain in a bioprocess, and

consequently, the selected strain might not be optimal from an economical/operational point of view. The solution to this problem can be addressed in two different ways. The first approach considers criteria related to bioprocess design in the early stages of strain design. This can be done by combining existing in silico strain optimization algorithms with dynamic flux balance analysis [23, 57, 58] to optimize yield (Y), titer (T), and productivity (P) in a balanced fashion. Zhuang [59] presented a Dynamic Strain Scanning Optimization (DySScO) strategy that uses this rationale to produce strains that balance the product yield, titer, and productivity. DySScO searches for a strain design that maximizes a user-defined metric of the form: $Z = f(Y, T, P)$. The second approach consists of decoupling the production of the target molecule from growth. The production process is thus divided into two phases. In the first phase, biomass is produced at a high rate and no production occurs. The second stage is characterized by low to no growth and production of the target chemical. The switching time from the growth phase to production has a high impact on the overall process performance.

Irrespective of the approach used, dynamic flux balance analysis (dFBA) has a central role in assessing and improving the performance of engineered strains in a bioreactor. dFBA combines both the process dynamics with the metabolic network, thus allowing the simulation of concentration profiles and flux distributions over time in the reactor and the cell, respectively. Shown here is the general procedure for performing a dFBA simulation with the COBRA Toolbox. Moreover, the utility of dFBA is illustrated with a case study.

1. Select a metabolic reconstruction and perform the necessary adjustment of the network (gene/reaction deletions, new pathways) in order to describe the metabolism of the strain studied. *See* **Note** 6.
2. Specify values for strain-specific parameters. This refers to substrate uptake and production rates, maximum growth rate, product inhibition, etc. These values can be taken from the literature or correspond to experimental measurements.
3. Define initial values for process-specific parameters. This refers to mode of operation (batch, fed batch), initial concentration of biomass and substrates, duration of the process, maximal reactor volume and feeding strategy (continuous substrate feeding, substrate pulses), time point of induction, etc.
4. Perform a dFBA simulation and determine values for yield, titer, and productivity.
5. Repeat steps 3 to 4 with a modified set of process-specific parameters until the desired performance for yield, titer, and productivity is reached. Alternatively, an optimization

algorithm can be used to find the set of optimal process-specific parameters that maximize yield, titer, and productivity.

The procedure explained above will, in the following, be used to study a hypothetical case for which experimental data is available. The production system consists of a strain carrying an inducible plasmid, which expresses heterologous enzymes necessary to synthetize some product R. In this hypothetical experiment, cells were cultivated until a defined optical density was reached, and then the synthesis of *R* was induced. The effect of varying the time point of induction on the yield, productivity, and titer will be analyzed. Figure 11 shows the experimentally obtained concentration profiles of biomass, glucose, and product inside the bioreactor over time. As can be concluded from the glucose concentration profile in Fig. 11 (circles, middle plot), the production of R occurs in a semibatch process in which glucose is added to the reactor in the form of two pulses over the course of the fermentation. The measured concentrations, shown as circles in Fig. 11, are used to determine glucose uptake and production rates before and after induction of the system. It is assumed that these parameters do not depend on the point of induction of the culture.

Figure 11a shows the consequences of varying the time point of induction for the plasmid-based system on the overall process performance. The dashed and dotted lines correspond to simulations performed with a modified time point of induction of $1.25 * t_{\mathrm{ind,exp}}$ or $0.5 * t_{\mathrm{ind,exp}}$, respectively. $t_{\mathrm{ind,exp}}$ refers to the experimental time point of induction and is used as a reference for the simulations. Increased product and biomass concentration is predicted by dFBA when the induction occurs at $1.25 * t_{\mathrm{ind,exp}}$. Interestingly, under this circumstance, the simulation also indicates that the actual fermentation setup would not support growth and production throughout the whole process time. The initial glucose concentration or the first glucose pulse has to be increased so that the process time is the same as in the experimental setup. The simulation also predicts the effect of a premature time point of induction. If the induction occurs at $1/2 * t_{\mathrm{ind,exp}}$, the biomass concentration remains comparatively low and the glucose concentration high. This in turn generates a lower end titer and productivity (right plot).

Figure 11b represents the behavior of the system which was simulated when the induction was made at $1.25 * t_{\mathrm{ind,exp}}$, and the first glucose pulse is sevenfold increased. This increase is realized to extend the process time of the simulated process thus allowing for a comparison with the experimental data. The improved feeding strategy leads to a twofold increase of the productivity and the end titer, as can be seen in Fig. 11b.

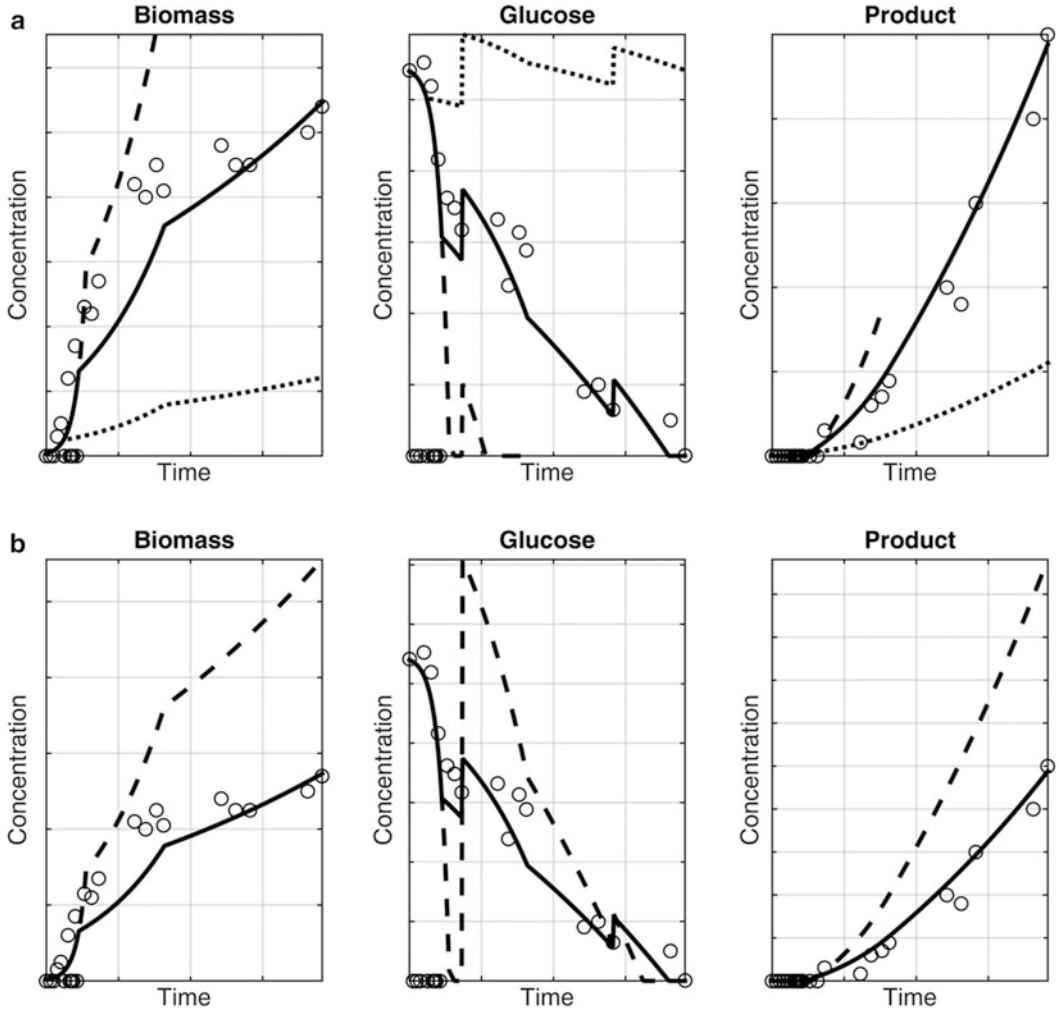


Fig. 11 Dynamic flux balance simulations for the production of a hypothetical target compound *R*. Biomass, glucose, and product concentration in the reactor over process time are shown. *Circles* (○) correspond to hypothetical measured concentrations. *Solid lines* represent simulated profiles using the experimental time point of induction, $t_{ind,exp}$. *Dashed lines* (-) were simulated with a time point of induction of $1.25^{*}t_{ind,exp}$ and *dotted* (•) lines with a time point of induction of $0.5^{*}t_{ind,exp}$. (**a**) The experimental feeding strategy is conserved for the simulations. (**b**) The first glucose pulse is increased sevenfold and the simulation for the process with an induction time point of $1.25^{*}t_{ind,exp}$ is performed again

4 Notes

1. Glpk is a free, widely used linear programming solver. However, its installation and use with the COBRA Toolbox can sometimes be difficult. Gurobi offers a good alternative when there is trouble with Glpk. At the homepage http://www.gurobi.com/, a free distribution can be downloaded for academic use.

2. The reaction `'SUCCt2_2'`, which transports succinate from the culture medium to the cytoplasm of the cell, has to be constrained to carry a reaction flux of zero. This prevents the occurrence of cycles when calculating the maximal theoretical succinate yield. These cycles lead to the reabsorption of the secreted succinate and thus generate an artificially high flux through the reaction `'SUCCt3'`. As a consequence, the calculated maximal theoretical yield has no biological meaning. This holds true for the calculation of the maximal theoretical yield for any target molecule using metabolic models. An easy way to verify the consistency is to calculate the carbon yield associated with the maximal yield being computed (molar or mass yield). Values for carbon yields greater than one are not consistent and need to be verified.
3. Many performance indices can be used in order to quantitatively assess the efficiency of a metabolic network with respect to the production of a target molecule. The most used performance index is the molar yield. Since the maximal value of this performance index depends on the substrate used and the target to be produced, it does not directly give an indication of the network efficiency. An alternative to the molar yield is the carbon yield. The carbon yield is related to the molar yield and has always, independent of the substrate used and target, a maximal value of 1. It provides therefore directly insight of the network performance and should be preferred if the network efficiency has an important role in the analysis being performed. If economic aspects should be considered, the maximal profit can be used. This performance index corresponds to the product of the molar yield and the market value of each component.
4. Because not only biomass but also intracellular fluxes vary with cellular behavior and time, for a stringent analysis, it is explicitly necessary to use corresponding rates that were measured simultaneously at the same time.
5. The flux distribution calculated for the wild-type or mutant strain using FBA is not unique in most cases. In order to compare the intracellular flux patterns of two strains, it is necessary to calculate the flux variability of the network [60]. This can be done by using the flux variability analysis (FVA) function of the COBRA Toolbox. Once the variability range for each reaction in the network has been calculated, it is recommended to limit the scope of the analysis to the reactions with a narrow or no variability range. Reactions with a broad variability range are often not essential for the network performance.

6. The suitability of a specific model that describes the behavior of a designed strain depends mainly on the accuracy of the assumptions made by the implemented model. For example, when modeling the metabolism of a strain that is used for heterologous protein production, a model that accounts for a fixed, comparatively low protein content will not be suitable for describing the behavior of the cell. In this case, a model that takes into account variable cell composition, such as a compartment model coupled with a metabolic model, will be more suitable for modeling the cell metabolism.

References

1. Trentacoste EM, Shrestha RP, Smith SR et al (2013) Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth. Proc Natl Acad Sci U S A 110:19748–19753. doi:10.1073/pnas.1309299110
2. Liang M-H, Jiang J-G (2013) Advancing oleaginous microorganisms to produce lipid via metabolic engineering technology. Prog Lipid Res 52:395–408. doi:10.1016/j.plipres.2013.05.002
3. Röling WFM, van Bodegom PM (2014) Toward quantitative understanding on microbial community structure and functioning: a modeling-centered approach using degradation of marine oil spills as example. Front Microbiol 5:125. doi:10.3389/fmicb.2014.00125
4. Sierra-García IN, Correa Alvarez J, de Vasconcellos SP et al (2014) New hydrocarbon degradation pathways in the microbial metagenome from Brazilian petroleum reservoirs. PLoS One 9, e90087. doi:10.1371/journal.pone.0090087
5. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. Nat Rev Microbiol 10:291–305. doi:10.1038/nrmicro2737
6. Lee JW, Kim TY, Jang Y-S et al (2011) Systems metabolic engineering for chemicals and materials. Trends Biotechnol 29:370–378. doi:10.1016/j.tibtech.2011.04.001
7. Ogata H, Goto S, Sato K et al (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:29–34. doi:10.1093/nar/27.1.29
8. Karp P (1996) EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. Nucleic Acids Res 24:32–39. doi:10.1093/nar/24.1.32
9. Schomburg I, Hofmann O, Baensch C et al (2000) Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. Gene Funct Dis 1:109–118. doi:10.1002/1438-826X(200010)1:3/4<109::AID-GNFD109>3.0.CO;2-O
10. Pirt SJ (1965) The maintenance energy of bacteria in growing cultures. Proc R Soc Lond Ser B Biol Sci 163:224–231
11. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121. doi:10.1038/nprot.2009.203
12. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320. doi:10.1038/msb.2009.77
13. Reed JL, Palsson BO (2003) Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. J Bacteriol 185:2692–2699. doi:10.1128/JB.185.9.2692-2699.2003
14. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 26:659–667. doi:10.1038/nbt1401
15. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biol 4:R54. doi:10.1186/gb-2003-4-9-r54
16. Feist AM, Henry CS, Reed JL et al (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121. doi:10.1038/msb4100155
17. Orth JD, Conrad TM, Na J et al (2011) A comprehensive genome-scale reconstruction

of *Escherichia coli* metabolism–2011. Mol Syst Biol 7:535. doi:10.1038/msb.2011.65

18. Covert MW, Famili I, Palsson BO (2003) Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? Biotechnol Bioeng 84:763–772. doi:10.1002/bit.10849
19. Hamilton JJ, Dwivedi V, Reed JL (2013) Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. Biophys J 105:512–522. doi:10.1016/j.bpj.2013.06.011
20. Beg QK, Vazquez A, Ernst J et al (2007) Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. Proc Natl Acad Sci U S A 104:12663–12668. doi:10.1073/pnas.0609845104
21. Covert MW, Schilling CH, Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. J Theor Biol 213:73–88. doi:10.1006/jtbi.2001.2405
22. Edwards JS, Ibarra RU, Palsson BO (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125–130. doi:10.1038/84379
23. Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. Appl Environ Microbiol 60:3724–3731
24. Pramanik J, Keasling JD (1997) Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng 56:398–421. doi:10.1002/(SICI)1097-0290(19971120)56:4<398::AID-BIT6>3.0.CO;2-J
25. Kremling A (2013) Systems biology: mathematical modeling and model analysis. CRC/Taylor & Francis, Boca Raton
26. Hyduke DR, Lewis NE, Palsson BØ (2013) Analysis of omics data with genome-scale models of metabolism. Mol Biosyst 9:167–174. doi:10.1039/c2mb25453k
27. Kim MK, Lun DS (2014) Methods for integration of transcriptomic data in genome-scale metabolic models. Comput Struct Biotechnol J 11:59–65. doi:10.1016/j.csbj.2014.08.009
28. Carrera J, Estrela R, Luo J et al (2014) An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *Escherichia coli*. Mol Syst Biol 10:735–735. doi:10.15252/msb.20145108
29. Murphy TA, Young JD (2013) ETA: robust software for determination of cell specific rates from extracellular time courses. Biotechnol Bioeng 110:1748–1758. doi:10.1002/bit.24836
30. Schellenberger J, Que R, Fleming RMT et al (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc 6:1290–1307. doi:10.1038/nprot.2011.308
31. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR (2013) COBRApy: COnstraints-based reconstruction and analysis for python. BMC Syst Biol 7:74. doi:10.1186/1752-0509-7-74
32. Klamt S, Saez-Rodriguez J, Gilles E (2007) Structural and functional analysis of cellular networks with Cell NetAnalyzer. BMC Syst Biol 1:2. doi:10.1186/1752-0509-1-2
33. Urbanczik R (2006) SNA – a toolbox for the stoichiometric analysis of metabolic networks. BMC Bioinformatics 7:129. doi:10.1186/1471-2105-7-129
34. Schwarz R, Musch P, von Kamp A et al (2005) YANA – a software tool for analyzing flux modes, gene-expression and enzyme activities. BMC Bioinformatics 6:135. doi:10.1186/1471-2105-6-135
35. Thiele I, Fleming RMT, Que R et al (2012) Multiscale modeling of metabolism and macromolecular synthesis in *E. coli* and its application to the evolution of codon usage. PLoS One 7:e45635. doi:10.1371/journal.pone.0045635
36. O'Brien EJ, Lerman JA, Chang RL et al (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol 9:693. doi:10.1038/msb.2013.52
37. Liu JK, O'Brien EJ, Lerman JA et al (2014) Reconstruction and modeling protein translocation and compartmentalization in *Escherichia* coli at the genome-scale. BMC Syst Biol 8:110. doi:10.1186/s12918-014-0110-6
38. Kremling A (2007) Comment on mathematical models which describe transcription and calculate the relationship between mRNA and protein expression ratio. Biotechnol Bioeng 96:815–819. doi:10.1002/bit.21065
39. Carta A (2014) Modelling, analysis and control for systems biology: application to bacterial growth models. Dissertation, University of Nice-Sophia Antipolis
40. Shen CR, Liao JC (2013) Synergy as design principle for metabolic engineering of 1-propanol production in *Escherichia coli*. Metab Eng 17:12–22. doi:10.1016/j.ymben.2013.01.008

41. Sánchez AM, Bennett GN, San K-Y (2005) Novel pathway engineering design of the anaerobic central metabolic pathway in *Escherichia coli* to increase succinate yield and productivity. Metab Eng 7:229–239. doi:10.1016/j.ymben.2005.03.001
42. Jantama K, Haupt MJ, Svoronos SA et al (2008) Combining metabolic engineering and metabolic evolution to develop nonrecombinant strains of C that produce succinate and malate. Biotechnol Bioeng 99:1140–1153. doi:10.1002/bit.21694
43. Zhang X, Jantama K, Moore JC et al (2009) Metabolic evolution of energy-conserving pathways for succinate production in *Escherichia coli*. Proc Natl Acad Sci 106:20180–20185. doi:10.1073/pnas.0905396106
44. Yang L, Cluett WR, Mahadevan R (2011) EMILiO: a fast algorithm for genome-scale strain design. Metab Eng 13:272–281. doi:10.1016/j.ymben.2011.03.002
45. Lin H, Bennett GN, San K-Y (2005) Genetic reconstruction of the aerobic central metabolism in *Escherichia coli* for the absolute aerobic production of succinate. Biotechnol Bioeng 89:148–156. doi:10.1002/bit.20298
46. Hoefel T, Faust G, Reinecke L et al (2012) Comparative reaction engineering studies for succinic acid production from sucrose by metabolically engineered *Escherichia coli* in fed-batch-operated stirred tank bioreactors. Biotechnol J 7:1277–1287. doi:10.1002/biot.201200046
47. Sánchez AM, Bennett GN, San K-Y (2006) Batch culture characterization and metabolic flux analysis of succinate-producing *Escherichia coli* strains. Metab Eng 8:209–226. doi:10.1016/j.ymben.2005.11.004
48. Wang W, Li Z, Xie J, Ye Q (2009) Production of succinate by a pflB ldhA double mutant of *Escherichia coli* overexpressing malate dehydrogenase. Bioprocess Biosyst Eng 32:737–745. doi:10.1007/s00449-009-0298-9
49. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84:647–657. doi:10.1002/bit.10803
50. Tepper N, Shlomi T (2010) Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics 26:536–543. doi:10.1093/bioinformatics/btp704
51. Patil KR, Rocha I, Förster J, Nielsen J (2005) Evolutionary programming as a platform for *in silico* metabolic engineering. BMC Bioinformatics 6:308. doi:10.1186/1471-2105-6-308
52. Lun DS, Rockwell G, Guido NJ et al (2009) Large-scale identification of genetic design strategies using local search. Mol Syst Biol 5:296. doi:10.1038/msb.2009.57
53. Ranganathan S, Suthers PF, Maranas CD (2010) OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS Comput Biol 6, e1000744. doi:10.1371/journal.pcbi.1000744
54. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. Mol Syst Biol 3:119. doi:10.1038/msb4100162
55. Schellenberger J, Park JO, Conrad TM, Palsson BØ (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11:213. doi:10.1186/1471-2105-11-213
56. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A 99:15112–15117. doi:10.1073/pnas.232349399
57. Mahadevan R, Edwards JS, Doyle FJ (2002) Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. Biophys J 83:1331–1340. doi:10.1016/S0006-3495(02)73903-9
58. Zhuang K, Ma E, Lovley DR, Mahadevan R (2012) The design of long-term effective uranium bioremediation strategy using a community metabolic model. Biotechnol Bioeng 109:2475–2483. doi:10.1002/bit.24528
59. Zhuang K, Yang L, Cluett WR, Mahadevan R (2013) Dynamic strain scanning optimization: an efficient strain design strategy for balanced yield, titer, and productivity. DySScO strategy for strain design. BMC Biotechnol 13:8. doi:10.1186/1472-6750-13-8
60. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264–276

Improving Biocontainment with Synthetic Biology: Beyond Physical Containment

Markus Schmidt and Lei Pei

Abstract

Genetically engineered organisms are per se subject to a biosafety risk assessment to define whether the resulting organism is safe for humans and the environment, either for contained use or environmental release. Contained use currently means physical containment and allows for a less strict assessment compared to environmental release. With developments in synthetic biology, we are currently witnessing the evolution of different forms of nonphysical containment enabled by sophisticated forms of genetic engineering, genome recoding, and xenobiology. Design and implementation of cells that use advanced suicide circuits, different genetic codes, alternative nucleic acids, amino acids, etc., will allow for a semantic or informational containment restricting and possibly eliminating horizontal gene flow with natural species. Here, we describe the scientific advances in this field and map the different approaches to design safe xeno-organisms. Finally, we address the questions that will have to be answered when semantic biocontainment systems become a reality.

Keywords: Biocontainment, Biosafety, Risk, Synthetic biology, Xenobiology

1 Introduction

As of 2015, practically all research and development projects in synthetic biology (SB) are done in contained facilities and at rather small scale. These activities carry only a very small probability regarding environmental release of genetically modified organisms (GMOs). Some applications suggested for SB, however, only make sense when there is a deliberate release into the environment (e.g., open pond microalgae production, bio-mining). But even those that entail the contained use of large-scale production facilities (e.g., production of bulk chemicals) face the risk of environmental escape of production organisms under real-world conditions [1–3].

To fully unlock the potential of SB, making SB applications safe regarding deliberate or accidental environmental release has become a hot and contested topic in research, policy, and public debate [4–11].

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 185–199,
DOI 10.1007/8623_2015_90, © Springer-Verlag Berlin Heidelberg 2015, Published online: 27 May 2015

Since the 1970s and 1980s, a number of ideas have been entertained and tested to provide a built-in safety system for GMOs, especially when physical containment alone is not enough [3, 11, 12]. Three major strategies have been applied for the design of microorganisms with genetic safeguards: (1) organisms with built-in auxotrophy (supplement is needed either to suppress a toxic gene production or to provide nutrition to survive), (2) induced lethality (kill switch), and (3) gene flow barriers (to integrate the circuits into the chromosome of the host or to include a killer gene which will be lethal to the new receipt hosts if a gene transfer event has happened) [3, 11].

In addition to that, several novel approaches have been proposed. One of them is an enhanced biocontainment system based on GeneGuard. This modular plasmid system consists of a conditional origin of replication that will limit the replication of the engineered plasmids in undesired hosts, a complementation of an introduced host auxotrophy that will replace the dependency of the antibiotic genes to maintain the plasmids in the host, and a toxin-antitoxin pair to prevent the plasmid spreading to other bacteria [13].

The other approach to build proper containment for the release of engineered or entirely synthetic microorganisms for bioremediation is to build genetic information exchange barriers by xenobiological approaches – using xeno-nucleic acids (XNAs) instead of DNA as information-bearing molecules, rewriting the genetic code to make it non-understandable by the existing gene expression machineries, and/or making growth dependent on xenobiotic chemicals [12].

It is known that all living organisms, from prokaryotic to eukaryotic species, have known strategies for genetic information exchange, which are critical for adaptation and evolution. One of these strategies is known as horizontal gene transfer (HGT). Taking the bacterial evolution, for example, the ability of bacteria to exploit new environment and mount response to new selective pressure is more likely due to new genes acquired by HGT than to "internal" mutations [14]. The mechanisms for HGT are mainly via conjugation, transduction, and natural transformation [15]. Mobile genetic elements, such as phages, transposons, and plasmids, play important roles in facilitating HGT. It is also known that these mobile genetic elements have played important roles in genetic engineering and now in SB research as well. Thus, the concern is that HGT can enable engineered organisms to evolve and circumvent current biocontainment designs, taking into account the high population number, mutation rate, and duplication time of microorganisms [16].

A handful approaches or concepts have been presented recently to improve nonphysical biocontainment strategies, including built-in genetic control circuits, genomically recoded organisms (GROs) that are engineered organisms with reassigned genetic code, and organisms with noncanonical biochemical building blocks (xenobiology). Here, we will review the recent scientific progress in

intrinsic biological containment, exploit the potential of the newly developed technologies for better biocontainment design, and discuss new research needs and challenges in the area of biocontainment of engineered organisms (Table 1).

2 Biocontainment Improved by Sophisticated Genetic Engineering Technologies

Recent progress in SB has made it possible to design more sophisticated genetic circuits to monitor environmental signals, broadening the applications of the engineered microorganisms from environmental biosensors to noninvasive diagnostic tools for human health. For example, *Escherichia coli* has been equipped with a genetic memory circuit capable of sensing, remembering, and reporting in the presence of antibiotic signals in the mammalian gut [17]. A recent article on the history of SB has reviewed the process on genetic circuits since 2000 [18]. Genetic circuits developed during the foundational years of contemporary SB (year 2000–2003) were those simple gene regulatory circuits mimicking minimalistic electric circuits, for example, a toggle switch [19] and a repressilator [20]. Campos wrote a historical overview of SB [21]. He noted that those SB systems developed during the intermediate years of SB (year 2004–2007) moved from simple to more complex systems, involving post-transcriptional [22] and translational controls [23], as well as circuits of multiple cellular pattern formations [24]. Those developed during the recent year of SB (year 2008–2013) were for more precisely controlled circuits, e.g., relaxation oscillator designed based on quantitative modeling other than those simple positive or negative feedback circuits [25].

So far, available intrinsic biocontainment systems based on built-in genetic designs are not considered to be failure-proof. While these systems could decrease the likelihood of HGT, they will not completely eliminate it [3, 11, 16]. Recent reviews showed that the frequencies of microbes that escaped engineered auxotrophy and lethality safeguard systems were from 5.00E-9 to 1.00E-4 [3], while the recommended limit was less than 1.00E-8 (from 1.00E-1 to 1.00E-8 observed among prokaryotes in the environment) [26]. Several concepts have been brought up to improve the biocontainment (i.e., to reduce the probability of HGT). These include, e.g., to design microbes with low environmental retention times, to remove mechanism of HGT (such as conjugation, transduction, or transformation), to design microbes with lower evolutional advantages (minimizing genes with marked selection advantage and/or including genes with marked selection disadvantage), and to design microbes with stacked containment strategies [11, 16]. Besides these approaches to improve biocontainment directly, other approaches might also help to improve the stability of the designed genetic circuits to reduce the uncertainty of the fate of the engineered organisms in the ecosystem. One observed failure of the

Table 1
Summary of existing and foreseeable biosafety mechanisms

	Conventional biocontainment (e.g., kill switch, auxotrophy)	Biocontainment with multiple conventional elements	Code reassignment	Genome recoding synthetic auxotrophic	Xenobiology
Escape frequency	1.00E-4 to 5.00E-9	NK	NK	1.00E-11 (below detection limit)	NK (possibly much below detection limit)
HGT	Possible	Possible	Unlikely	Unlikely	NK
Possible resistant mechanisms	Mutations; auxotrophy complemented by natural supplements or symbiotic relationships	Mutations of the genes encoding multiple-layered safety guards	NK	Possible resistance mechanisms of the synthetic auxotrophic strains have been overcome by further modifications	NK

NK stands for not known

current genetic containment is due to the weak stability of the safeguard structures [3]. To tackle this issue, rational design of genetic circuits with improved evolutionary stability has been studied. One example was to combine the target genes (genes of interest) with the essential genes under the regulation of the bidirectional promoters (forward and backward). While a target gene (e.g., green fluorescent protein) coupled to the essential gene (e.g., Kanamycin resistant gene) in a bidirectional promoter circuit, the evolutionary half time of it showed 4–10 times increase in *E. coli* [27].

The other risk of failure of the biocontainment might be due to the plasmids used to construct the safeguard. The usage of plasmids has facilitated the genetic engineering yet also increased the likelihood of gene transfer. The recombinant plasmids might be transferred to other organisms through the active manner via conjugation or the passive manner via uptake of those released from the dead cells [3]. The progress of genome editing might provide solutions to overcome or significantly reduce the risk of gene transfer mediated by plasmids. The recently developed genome-editing technologies are mediated by Zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and the emergence of clustered regulatory interspaced short palindromic repeat (CRISPR)/Cas-based RNA-guided DNA endonucleases. These chimeric nucleases enable a broad range of genetic modifications in the target organisms [28–31]. Meanwhile, the new progress of whole genome synthesis, ranging from bacterial genomes, yeast genomes, to mammalian mitochondrial genomes, might help to rationally design organisms equipped with all possible semantic biocontainment components to maximize the genetic flow safeguard system [32–34]. As of 2014, this idea remains a theoretical concept since real-world (contained) experiments testing the efficacy are not (yet) taking place.

In addition, more knowledge is needed to better estimate the frequency of HGT of constructed genetic circuits. Statistical approaches have been proposed to access the probability of HGT events in microbial populations of four different case scenarios [15]. These approaches developed based on the knowledge of the rates of the processes (yet independent of the mechanism of HGT) would be applicable to HGT events occurring between unrelated species. It is clear, however, that a theoretical calculation alone will not be enough to assess its real-world practicability. Practical experiments need to follow, such as HGT experiments on filter or soil and water microcosms.

3 Biocontainment Approaches Based on Codon Reassignment

Biocontainment can also be enhanced by changing the biological semantics of engineered organisms. With semantics, we mean the genetic language used by cells, i.e., the cell uses four different

bases as letters, combining them to three letter words (triplets) which are arranged to sentences that gain meaning via their translation into amino acids and proteins. The term "genetic code" describes how the 64 triplets are translated to the 20 (up to 22) amino acids. The translational fidelity in the proper codon assignment for 20 canonical amino acids (AAs) has played an important role in the high precision, robustness, and stability of the translation from the genetic to the protein alphabet. This set of code is shared by almost all organisms, which is why it is called the "standard" genetic code. Besides the standard code, science currently knows 24 other naturally occurring codes, such as the one used in vertebrate mitochondrial DNA. HGT only works in nature because the source and the receptor organism both use the same code; in the case where they don't share the same code, the foreign genetic information would make no sense to the receiving organism.

Codon reassignment techniques have been developed to engineer microorganisms that also accept noncanonical amino acids (ncAAs) in the translational machinery [35–38]. Restricted selection pressure can lead to the reassignment of certain genetic codons to some ncAAs.

To reassign codons on demand will require different approaches other than harnessing the natural existing mechanisms. One approach to achieve codon reassignment is to eliminate the aminoacylation proofreading step. Another approach is to exploit the broad ribosome substrate specificity, e.g., to change or relax the substrate specificity of the aminoacyl-tRNA synthetases (aaRS) [35]. More than 70 ncAAs have been genetically encoded by reassignments to blank codons in *E. coli*, yeast, and mammalian cells [39].

One challenging research endeavor is to reassign the amber stop codon UAG to a desired ncAA. Such an approach targets the stop codon, the so-called stop codon suppression (SCS) methodologies using a heterologous orthogonal aaRS-tRNA pair to incorporate an ncAA in response to a stop codon [36, 39].

Even further goes the scrambling of the genetic code by switching from 3 to 4 bases as "genetic words." By introducing an orthogonal ribosome (ribo-Q1), a serial of quadruplet codons can be assigned to encode ncAAs. The approach based on combining synthetases-tRNA pair and ribo-Q1 can encode 256 blank codons in theory [40]. An additional approach to incorporate ncAAs in vivo is to synthesize ncAA by pyridoxal 5′-phosphate (PLP)-dependent enzymes [38]. PLP-dependent enzymes are known to catalyze several essential chemical reactions, such as transamination, decarboxylation, racemization, carbon–carbon bond cleavage, and formation. Combining a PLP-dependent enzyme metabolic biosynthesis of ncAAs could set the microorganisms with codon reassignment free from the dependence on specific supplement on additional chemicals to survive. An engineering approach to couple

the orthogonal chemistries with artificial metabolism was achieved in a methionine-auxotroph *E. coli* strain to directly incorporate ncAA into a recombinant protein (barstar) [41]. The intracellular biosynthesis metabolic pathway was engineered to produce L-azidohomoalanine from *O*-acetyl-L-homoserine and NaN^3. The direct incorporation of this ncAA into recombinant protein barstar was achieved by the production of the ncAA by exploiting the broad specificity of recombinant pyridoxal phosphate-dependent *O*-acetylhomoserine sulfhydrylase from *Corynebacterium glutamicum* and AUG codon reassignment to incorporate L-azidohomoalanine in place of L-methionine [41]. This approach showed that ncAA could be produced by the engineered intracellular biosynthesis using common fermentable sources, paving way to develop novel approaches on using external food sources to contain the organisms engineered based on codon reassignment. The value of the codon reassignment for enhanced biocontainment lies more on preventing the built-in heterogeneous genes (encoding functional proteins) in the engineered organisms from being correctly expressed in other organisms. So even though when HGT does happen between the engineered and another species, the other species is unable to translate the proteins [42]. So far, codon reassignment has just left the proof of principle phase. Real-world test regarding its value to improve biocontainment has not been made.

A natural example of a mild semantic biocontainment has been described in a recent research paper that showed that UGA (normally a stop codon) is an additional glycine codon in uncultured SR1 bacteria from the human oral microbiota [43]. It is known that many human cohabiting microbes from phylum SR1 are difficult to cultivate and are only identified by small subunit rRNA sequences. Single-cell genome sequence on one such taxon (SR1-OR1) from a healthy oral sample revealed that this SR1 bacteria use a unique genetic code, where the UGA codon is not a stop codon but in equilibrium with the canonical GGN glycine codons. It seems that UGA codon reassignment prevented the SR1 genes from being translatable by other bacteria. That means the unique codon reassignment strategy helps the SR1 bacteria to keep the advantageous genetic information among their own species, not sharing it with other microbes via HGT in the human microbiota. This provides a proof of principle that the codon reassignment can help to contain the genes of interest within the engineered ones but not the native ones.

Another natural codon reassignment on canonical genetic code (CUU and CUA sense codons to alanine instead of leucine in the standard code) has been discovered in the mitochondria of *Ashbya* (*Eremothecium*) *gossypii*, a filamentous-growing plant pathogen related to yeast Saccharomycetaceae [44]. Natural codon reassignments, although not universal, can be found in the genomes (11 events) and mitochondria (16 events), indicating

the organisms with codon reassignment can survive well in the natural setting.

The 25 known codes are, however, an insignificant share of the combinatorial explosion of theoretical possible genetic codes. A simple calculation shows that there are more than 10^{71} possible genetic codes.Semantic biocontainment tries to achieve a Babylonian diversity of codes to render HGT meaningless.

4 Biocontainment Approaches Based on Genome Recoding

The (almost) universal genetic code shared by most microbes allows the expression of heterogeneous genes in the engineered microorganisms. However, this common system also permits organisms to exchange genes through HGT, posing a challenge to genetic containment for environmental applications.

Recently, GROs have been constructed to build microbes with improved physiological properties (such as viral resistance and efficient incorporation of the nonstandard amino acids, nsAAs) and better biosafety profiles [45–48]. An in vivo genome-editing approach was applied to replace all known 321 UAG stop codons in *E. coli* MG1655 with synonymous UAA codons. In this GRO strain, the UAG termination (release factor 1, RF1) was further eliminated, which allowed the reintroduced UAG to code for the ncAAs (e.g. *p*-acetylphenylalanine, pAcF) [48]. GROs might expand the chemical capabilities of the engineered microbes and isolate them better from nature. A biocontainment approach based on GROs with recoding on stop codon might help to prevent GROs gain heterogeneous genes from natural organisms due to the read-through on the translation terminators resulting in mistranslation of the foreign genes.

Several challenges have been raised for the genome recoding due to the fact that codon usage can strongly affect gene regulation and translation; in turn, genome-wide recoding on essential genes remains unexplored until the limitation on genetic recoding in essential genes has been studied [47]. Codon reassignments on "13 forbidden codes" were done in 42 highly expressed essential genes in 80 *E. coli* strains, of which 41 were essential ribosomal protein-coding genes and the *prfB* (the gene encoding RF2). The strains with recoded genes exhibited the broadest range of fitness defects (measured by doubling time) and indicated that although individual gene recoding was feasible, pooling together several recoded genes into one genome could "lead to fitness impairment." This requires to be explored further when a genome recoding involves essential genes. In 2015, engineered microorganisms dependent on synthetic amino acids to survive were developed by two US research groups [49, 50]. Both engineered microbes developed by these two groups

were obtained by further modifications on a recoded *E. coli* strain C321. Δ A which was developed earlier [48]. Mandell et al. engineered the strain further by recoding the UAG for L-4,4′-biphenylalmine (bipA), which eventually resulted in a triple synthetic auxotrophic strain with two essential proteins (adenylate kinase and tyrosyl-tRNA synthetase) that incorporated bipA and the re-engineered bipA aaRS that required bipA for folding. The resulting synthetic auxotrophic strains proved unable to metabolically bypass the biocontainment mechanisms by compounds found in the natural environment, while they showed low escape rates and were resistant to HGT [49]. Instead of incorporating bipA into essential proteins, Rovner et al. recoded UAG in C321. Δ A for three synthetic amino acids: *p*-acetyl-L-phenylalanine (pAcF), *p*-iodo-L-phenylalanine (pIF), or *p*-azido-L-phenylalanine (pAzF). The multiple-layered safeguard engineered synthetic auxotrophic strains obtained were the strains incorporating pAcF in the functional sites of three essential proteins, while the genes encoding tyrosine tRNAs (*tyrT* and *tyrV*) were deleted [50]. This novel GRO strain showed improved containment profiles: low escape rate (undetectable escape frequencies upon culturing 10^{11} cells on solid media for 7 days or in liquid media for 20 days) and resistance to HGT [50]. These two studies showed that better containment mechanisms could be developed based on synthetic auxotrophic. They also demonstrated, however, that the currently available metrics to measure the degree of safety are limited. Mandell et al. [49] stated: "Our results demonstrate that mutational escape frequency under laboratory growth conditions is a necessary but insufficient metric to evaluate biocontainment strategies." Thus, new standards and metrics are needed to quantitatively define how safe these new strains really are.

Although the genome recoding in eukaryotic cells is less studied, such concept has been applied as well to on-going research. Taking one example from the *Sc2.0* project [51], to enhance the genetic flexibility of the synthetic chromosome arms that could function in yeast, the elimination of the TAG stop codons of the right arm of *Saccharomyces cerevisiae* chromosome IX (IXR) was done by recoding them to TAA, thus reserving one codon for future expansion of the genetic code, e.g., to add ncAA [45]. This approach could also serve as a future mechanism for genetic isolation and an additional level of containment control over the synthetic yeast. In addition, an error-prone orthogonal DNA replication system has been developed in *S. cerevisiae*, a system consisting of an orthogonal plasmid- polymerase pair. This plasmid-polymerase-based system could allow increased substantial mutation rates of the plasmid (400-fold greater than the host genome), while the mutation rates of host genome were not affected. This system can serve as a platform for in vivo continuous evolution [52].

5 Directed Evolution

Although genome recoding may require sophisticated engineering approaches, there are alternative approaches to rewrite the genome with a different genetic makeup. One of these attempts was achieved by artificially evolving the microbial genome in a turbidostat (the dynamic sister of the chemostat) to incorporate the non-canonical thymidine analogue 5-chlorouracil. This approach has been applied to generate an *E. coli* strain that contained 90% chlorodeoxyuridine and 10% thymidine [53]. Chemically modified organisms might lead the way toward a new type of genetic firewall [54, 55].

Due to the genome dynamics and the fact that reversion is an evolutionary process, one concern about the semantic biocontainment efficiency of genome recoding is the slow fitness recovery in the modified genomes. Not much research has been done on this aspect. However, knowledge on the codon-modified viral genome might provide some hints. Synonymous codon modification of viral genome has been carried out in many viruses and considered as an effective approach to develop attenuated vaccines [56–58]. This approach for attenuating viruses for vaccine was based on a different principle used in genome recoding though: changing a large number of codons within the viral genome but not changing the protein sequence. The key of this approach is to replace the wild-type codons with designed codons of those sequences that impair replication and/or expression. The slow fitness recovery of the codon-attenuated viruses to high fitness or even to high virulence has been studied in a codon-modified bacterial virus T7 [59]. Results on evaluations on the fitness of the engineered viruses, the evolution of the sequence, and the fitness effects of the changes supported "the premise that codon-modified viruses recover fitness slowly, although the evolution is substantially more rapid than expected from the design principle." Besides the modified viral genome, the genome evolution and adaptation of a bacterial strain were also studied. Genomes of 40,000 generations from a common laboratory *E. coli* strain were sequenced [60]. It showed that genomic evolution was constant for the first 20,000 generations despite the sharp decline of adaptation. Microorganisms of these first 20,000 generations showed low mutation rate, and those mutations were neutral. The studies on the sequences from the 40,000 generations showed that synonymous change rates were lower than the first 20,000 generations, while the mutants found in the 40,000 genome were skewed toward AT to GC transversions. The continuous investigation on the evolution and adaptation on the genomes should also be applied to the recoded genomes in the future.

6 Biocontainment Approach Based on Xenobiology

Xenobiology (XB) is the design, engineering, and production of biological systems with noncanonical biochemistries and/or alternative genetic codes. XB is a subfield of synthetic biology, and in addition to genetic code engineering and the use of ncAAs, it also covers XNAs, expanded genetic alphabet with alternative base pairs, and the use of novel polymerases and ribosomes.

XNAs: The chemical backbone of DNA and RNA is desoxyribose and ribose, respectively, and appears to be highly conserved biochemical structures in nature [61, 62]. When another chemical structure is used as a base-carrying backbone, the abbreviation of the resulting nucleic acid changes, e.g., to HNA (hexose), CeNA (cyclohexenyl), or TNA (threose) [63–65]. The collective term for all nucleic acids that are not DNA or RNA is thus XNA, where the X stands for xeno (Greek for "foreign or unknown") [54, 66].

Expanded genetic alphabet or alternative base pairs: The two natural base pairs in DNA are A-T (A-U in RNA) and C-G. From a chemical point of view, these base pairs match because their chemical architecture and the number of hydrogen bonds fit together (A-T has two and C-G has three hydrogen bonds). C and T are pyrimidines, while A and G are purines. Additional base pairs can be synthesized and incorporated into DNA (or XNA). In case the aim is to extend the genetic alphabet, the new base pairs (or the new base in case it is self-pairing) need to match each other with high accuracy and discriminate against other existing bases to maintain information storage capabilities [67]. For each added base pair, the genetic alphabet grows by 2; in the special case of a self-pairing base, it would grow by 1. For example, in 2014, researchers announced that they had successfully introduced two new artificial nucleotides into bacterial DNA, alongside the four naturally occurring nucleotides, and, by including individual artificial nucleotides in the culture media, were able to retain this new base pair for several days [68].

Novel polymerases and ribosomes: In most cases, natural polymerases and ribosomes (and other nucleic acid-interacting proteins) do not work on XNAs and nucleic acids with expanded alphabets. To allow for replication, transcription, and translation, the nucleic acid cell machinery has to be adapted to operate on these novel nucleic acids [69].

Combining these three types of alterations in living organisms will lead to different degrees of xeno-organisms that will have less and less in common with their naturally evolved counterparts. Xeno-organisms will not be able to read or translate the genetic

information into tangible products for the cell, thus rendering the original information useless and out of reach. HGT will thus be severely limited, if not entirely impossible.

7 Conclusion and Future Challenges

SB and especially xenobiology will add a powerful set of tools and methods to improve biocontainment beyond physical restrictions. Semantic or informational containment refers to the use of different biological languages (genetic codes) or building blocks (e.g., nucleic acids) as an additional layer of safety for upcoming biotechnological devises and systems. The proof of principle for all constituting aspects of xenobiology has been given in recent years. The coming years will show how easy or difficult it will be to integrate the different systems into one organism and "move it as far away" from natural cells as possible. Soon, we will have to ask ourselves: should these novel species be treated the same or differently from currently genetically engineered organisms, and at what point will the semantic biocontainment be "strong enough" to use it even without physical containment [70]? What kind of metric should be used to measure the semantic distance between natural and different types of xeno-organisms and what metric to measure the biosafety increase? And what, finally, will be suitable areas of application for xeno-organisms? These and more questions will have to be answered as xenobiology develops from the stage of "proof of principle" to become a mainstream methodology enabling applications for the real world.

Acknowledgments

Author acknowledges the financial support of the EC-FP7 project METACODE (EC Grant No. 289572), and MS acknowledges the financial support of the EC-FP7 project ST-FLOW (EC Grant No. 289326).

References

1. Khalil AS, Collins JJ (2010) Synthetic biology: applications come of age. Nat Rev Genet 11 (5):367–379
2. Armstrong R, Schmidt M, Bedau M (2012) Other developments in synthetic biology. In: Schmidt M (ed) Synthetic biology industrial and environmental applications. Wiley, Weinheim
3. Moe-Behrens GH, Davis R, Haynes KA (2013) Preparing synthetic biology for the world. Front Microbiol 4:5
4. DFG, acatech and Lepoldina (2009) Synthetic biology: positions. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2009/stellungnahme_synthetische_biologie.pdf
5. EGE (2009) Ethics of synthetic biology. http://ec.europa.eu/bepa/european-group-ethics/docs/opinion25_en.pdf
6. Gaisser S, Reiss T, Lunkes A, Muller K, Bernauer H (2009) Making the most of synthetic biology. Strategies for synthetic biology

development in Europe. EMBO Rep 10(Suppl 1):S5–S8
7. Schmidt M, Ganguli-Mitra A, Torgersen H, Kelle A, Deplazes A, Biller-Andorno N (2009) A priority paper for the societal and ethical aspects of synthetic biology. Syst Synth Biol 3(1–4):3–7
8. The Royal Academy of Engineering (2009) Synthetic biology: scope, applications and implications. The Royal Academy of Engineering, London
9. Bubela T, Hagen G, Einsiedel E (2012) Synthetic biology confronts publics and policy makers: challenges for communication, regulation and commercialization. Trends Biotechnol 30(3):132–137
10. Torgersen H, Schmidt M (2013) Frames and comparators: how might a debate on synthetic biology evolve? Futures 48(100):44–54
11. Wright O, Stan GB, Ellis T (2013) Building-in biosafety for synthetic biology. Microbiology 159(Pt 7):1221–1235
12. Schmidt M, de Lorenzo V (2012) Synthetic constructs in/for the environment: managing the interplay between natural and engineered Biology. FEBS Lett 586(15):2199–2206
13. Wright O, Delmans M, Stan G-B, Ellis T (2014) GeneGuard: a modular plasmid system designed for biosafety. ACS Synth Biol 4:307–316
14. Davison J (1999) Genetic exchange between bacteria in the environment. Plasmid 42 (2):73–91
15. Townsend JP, Bohn T, Nielsen KM (2012) Assessing the probability of detection of horizontal gene transfer events in bacterial populations. Front Microbiol 3:27
16. Marris C, Jefferson C (2013) Workshop on "Synthetic biology: containment and release of engineered micro-organisms" held on 29 April 2013 at King's College London: Summary of Discussions. https://kclpure.kcl.ac.uk/portal/en/publications/workshop-on-synthetic-biology-containment-and-release-of-engineered-microorganisms-held-on-29-april-2013-at-kings-college-london%28df5f0ce4-61a9-4067-9705-1851926aa2a2%29/export.html
17. Kotula JW, Kerns SJ, Shaket LA, Siraj L, Collins JJ, Way JC, Silver PA (2014) Programmable bacteria detect and record an environmental signal in the mammalian gut. Proc Natl Acad Sci U S A 111:4838–4843
18. Cameron DE, Bashor CJ, Collins JJ (2014) A brief history of synthetic biology. Nat Rev Microbiol 12:381–390
19. Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in *Escherichia coli*. Nature 403(6767):339–342
20. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403(6767):335–338
21. Campos L (2009) That was the synthetic biology that was. In: Schmidt M, Kelle A, Ganguli-Mitra A, de Vriend H (eds) Synthetic biology: the technoscience and its societal consequences. Springer, Dordrecht, pp 5–21
22. Isaacs FJ, Dwyer DJ, Ding C, Pervouchine DD, Cantor CR, Collins JJ (2004) Engineered riboregulators enable post-transcriptional control of gene expression. Nat Biotechnol 22 (7):841–847
23. Bayer TS, Smolke CD (2005) Programmable ligand-controlled riboregulators of eukaryotic gene expression. Nat Biotechnol 23 (3):337–343
24. Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R (2005) A synthetic multicellular system for programmed pattern formation. Nature 434(7037):1130–1134
25. Stricker J, Cookson S, Bennett MR, Mather WH, Tsimring LS, Hasty J (2008) A fast, robust and tunable synthetic gene oscillator. Nature 456(7221):516–519
26. Brigulla M, Wackernagel W (2010) Molecular aspects of gene transfer and foreign DNA acquisition in prokaryotes with regard to safety issues. Appl Microbiol Biotechnol 86 (4):1027–1041
27. Yang S, Sleight SC, Sauro HM (2013) Rationally designed bidirectional promoter improves the evolutionary stability of synthetic genetic circuits. Nucleic Acids Res 41(1):e33
28. Carroll D (2011) Zinc-finger nucleases: a panoramic view. Curr Gene Ther 11(1):2–10
29. Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE, Leung E, Hinkley SJ, Dulay GP, Hua KL, Ankoudinova I, Cost GJ, Urnov FD, Zhang HS, Holmes MC, Zhang L, Gregory PD, Rebar EJ (2011) A TALE nuclease architecture for efficient genome editing. Nat Biotechnol 29(2):143–148
30. Gaj T, Gersbach CA, Barbas CF 3rd (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31(7):397–405
31. Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 32(4):347–355
32. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C,

Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova EA, Young L, Qi ZQ, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA 3rd, Smith HO, Venter JC (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329(5987):52–56

33. 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
34. Pennisi E (2014) Building the ultimate yeast genome. Science 343:1426–1429
35. Budisa N, Minks C, Alefelder S, Wenger W, Dong F, Moroder L, Huber R (1999) Toward the experimental codon reassignment in vivo: protein building with an expanded amino acid repertoire. FASEB J 13(1):41–51
36. Hoesl MG, Budisa N (2012) Recent advances in genetic code engineering in *Escherichia coli*. Curr Opin Biotechnol 23(5):751–757
37. Budisa N (2013) Expanded genetic code for the engineering of ribosomally synthetized and post-translationally modified peptide natural products (RiPPs). Curr Opin Biotechnol 24 (4):591–598
38. di Salvo ML, Budisa N, Contestabile R (2013) PLP-dependent Enzymes: a powerful tool for metabolic synthesis of non-canonical amino acids. In: Beilstein Bozen symposium on molecular engineering and control. Beilstein Institute, Prien, pp 27–66
39. Liu CC, Schultz PG (2010) Adding new chemistries to the genetic code. Annu Rev Biochem 79:413–444
40. Neumann H, Wang K, Davis L, Garcia-Alai M, Chin JW (2010) Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. Nature 464 (7287):441–444
41. Ma Y, Biava H, Contestabile R, Budisa N, di Salvo ML (2014) Coupling bioorthogonal chemistries with artificial metabolism: intracellular biosynthesis of azidohomoalanine and its incorporation into recombinant proteins. Molecules 19(1):1004–1022
42. Doering V (2007) Sense codon reassignment as means of synthesizing safe genetically engineered microorganism. SB3.0, Zurich
43. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, Soll D, Podar M (2013) UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proc Natl Acad Sci U S A 110(14):5540–5545
44. Ling J, Daoud R, Lajoie MJ, Church GM, Soll D, Lang BF (2014) Natural reassignment of CUU and CUA sense codons to alanine in Ashbya mitochondria. Nucleic Acids Res 42(1):499–508
45. Dymond JS, Richardson SM, Coombes CE, Babatz T, Muller H, Annaluru N, Blake WJ, Schwerzmann JW, Dai J, Lindstrom DL, Boeke AC, Gottschling DE, Chandrasegaran S, Bader JS, Boeke JD (2011) Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. Nature 477(7365):471–476
46. Isaacs FJ, Carr PA, Wang HH, Lajoie MJ, Sterling B, Kraal L, Tolonen AC, Gianoulis TA, Goodman DB, Reppas NB, Emig CJ, Bang D, Hwang SJ, Jewett MC, Jacobson JM, Church GM (2011) Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. Science 333 (6040):348–353
47. Lajoie MJ, Kosuri S, Mosberg JA, Gregg CJ, Zhang D, Church GM (2013) Probing the limits of genetic recoding in essential genes. Science 342(6156):361–363
48. Lajoie MJ, Rovner AJ, Goodman DB, Aerni HR, 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
49. Mandell DJ, Lajoie MJ, Mee MT, Takeuchi R, Kuznetsov G, Norville JE, Gregg CJ, Stoddard BL, Church GM (2015) Biocontainment of genetically modified organisms by synthetic protein design. Nature 518:55–60
50. Rovner AJ, Haimovich AD, Katz SR, Li Z, Grome MW, Gassaway BM, Amiram M, Patel JR, Gallagher RR, Rinehart J, Isaacs FJ (2015) Recoded organisms engineered to depend on synthetic amino acids. Nature 518:89–93
51. Sc2.0. "Synthetic Yeast 2.0". http://biostudio.bme.jhu.edu/sc2/?page_id=63

52. Ravikumar A, Arrieta A, Liu CC (2014) An orthogonal DNA replication system in yeast. Nat Chem Biol 10:175–177
53. Marliere P, Patrouix J, Doring V, Herdewijn P, Tricot S, Cruveiller S, Bouzon M, Mutzel R (2011) Chemical evolution of a bacterium's genome. Angew Chem Int Ed Engl 50 (31):7109–7114
54. Schmidt M (2010) Xenobiology: a new form of life as the ultimate biosafety tool. Bioessays 32 (4):322–331
55. Acevedo-Rocha CG, Budisa N (2011) On the road towards chemically modified organisms endowed with a genetic firewall. Angew Chem Int Ed Engl 50(31):6960–6962
56. Burns CC, Shaw J, Campagnoli R, Jorba J, Vincent A, Quay J, Kew O (2006) Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. J Virol 80(7):3259–3272
57. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S (2008) Virus attenuation by genome-scale changes in codon pair bias. Science 320(5884):1784–1787
58. Mueller S, Coleman JR, Papamichail D, Ward CB, Nimnual A, Futcher B, Skiena S, Wimmer E (2010) Live attenuated influenza virus vaccines by computer-aided rational design. Nat Biotechnol 28(7):723–726
59. Bull JJ, Molineux IJ, Wilke CO (2012) Slow fitness recovery in a codon-modified viral genome. Mol Biol Evol 29(10):2997–3004
60. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461(7268):1243–1247
61. Eschenmoser A (1999) Chemical etiology of nucleic acid structure. Science 284 (5423):2118–2124
62. Pace NR (2001) The universal nature of biochemistry. Proc Natl Acad Sci U S A 98 (3):805–808
63. Chaput JC, Ichida JK, Szostak JW (2003) DNA polymerase-mediated DNA synthesis on a TNA template. J Am Chem Soc 125 (4):856–857
64. Ichida JK, Horhota A, Zou K, McLaughlin LW, Szostak JW (2005) High fidelity TNA synthesis by Therminator polymerase. Nucleic Acids Res 33(16):5219–5225
65. Kempeneers V, Renders M, Froeyen M, Herdewijn P (2005) Investigation of the DNA-dependent cyclohexenyl nucleic acid polymerization and the cyclohexenyl nucleic acid-dependent DNA polymerization. Nucleic Acids Res 33(12):3828–3836
66. Marliere P (2009) The farther, the safer: a manifesto for securely navigating synthetic species away from the old living world. Syst Synth Biol 3(1–4):77–84
67. Yang Z, Hutter D, Sheng P, Sismour AM, Benner SA (2006) Artificially expanded genetic information system: a new base pair with an alternative hydrogen bonding pattern. Nucleic Acids Res 34(21):6095–6101
68. Malyshev DA, Dhami K, Lavergne T, Chen T, Dai N, Foster JM, Correa IR, Romesberg FE (2014) A semi-synthetic organism with an expanded genetic alphabet. Nature 509:385–388
69. Pinheiro VB, Taylor AI, Cozens C, Abramov M, Renders M, Zhang S, Chaput JC, Wengel J, Peak-Chew SY, McLaughlin SH, Herdewijn P, Holliger P (2012) Synthetic genetic polymers capable of heredity and evolution. Science 336 (6079):341–344
70. SCHER, SCENIHR, SCCS (2014) Preliminary opinion on synthetic biology I definition. http://ec.europa.eu/health/scientific_committees/emerging/docs/scenihr_o_044.pdf