1 Molecular Biotechnology: From DNA Sequence to Therapeutic Protein

Ronald S. Oosting

INTRODUCTION

Proteins are already used for more than 100 years to treat or prevent diseases in humans. It started in the early 1890s with "serum therapy" for the treatment of diphtheria and tetanus by Emile von Behring and others. The antiserum was obtained from immunized rabbits and horses. Behring received the Nobel Prize for Medicine in 1901 for this pioneering work on passive immunization. A next big step in the development of therapeutic proteins was the use of purified insulin isolated from pig or cow pancreas for the treatment of diabetes type I in the early 1920s by Banting and Best (in 1923 Banting received the Nobel Prize for this work). Soon after the discovery of insulin, the pharmaceutical company Eli Lilly started large-scale production of the pancreatic extracts for the treatment of diabetes. Within 3 years after the start of the experiments by Banting, already enough animal-derived insulin was produced to supply the entire North American continent. Compare this to the present average time-to-market of a new drug (from discovery to approval) of 13.5 years (Paul et al. 2010).

Thanks to advances in biotechnology (e.g., recombinant DNA technology, hybridoma technology), we have moved almost entirely away from animal-derived proteins to proteins with the complete human amino acid sequence.

Such therapeutic human proteins are less likely to cause side effects and to elicit immune responses. Banting and Best were very lucky. They had no idea about possible sequence or structural differences between human and porcine/bovine insulin. Nowadays, we know that porcine insulin differs only with one amino acid from the human sequence and bovine insulin differs by three amino acids (see Fig. 1.1). Thanks to this high degree of sequence conservation, porcine/bovine insulin can be used to treat human patients. In 1982, human insulin became the first recombinant human protein approved for sale in the USA (also produced by Eli Lilly) (cf. Chap. 12). Since then a large number of biopharmaceuticals have been developed. There are now almost 200 human proteins marketed for a wide range of therapeutic areas.

PHARMACEUTICAL BIOTECHNOLOGY, WHY THIS BOOK, WHY THIS CHAPTER?

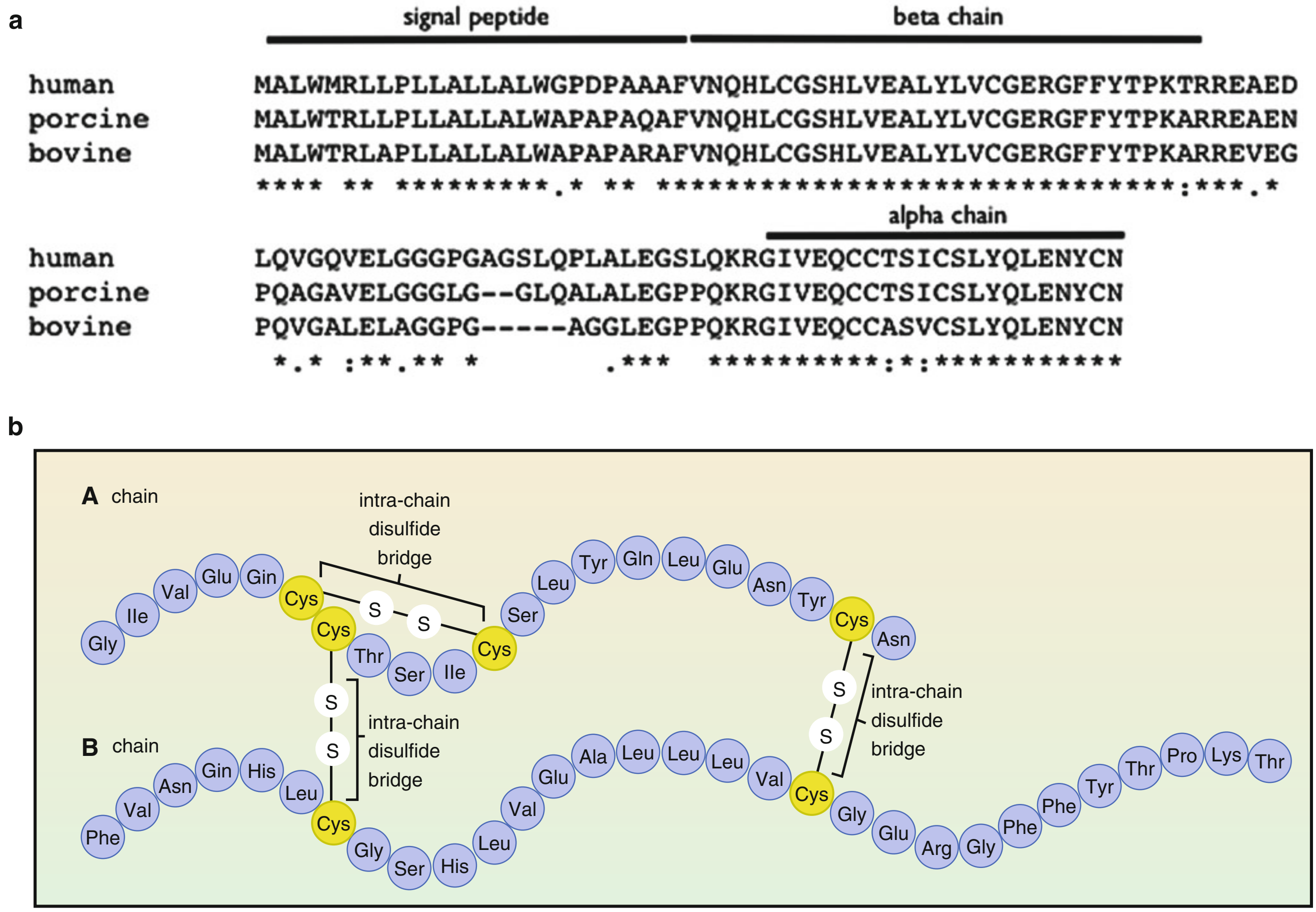
In this book we define pharmaceutical biotechnology as all technologies needed to produce biopharmaceuticals (other than (nongenetically modified) animal- or human blood-derived medicines). Attention is paid both to these technologies and the products thereof. Biotechnology makes use of findings from various research areas, such as molecular biology, biochemistry, cell biology, genetics, bioinformatics, microbiology, bioprocess engineering, and separation technologies. Progress in these fields has been and will remain a major driver for the development of new biopharmaceuticals. Biopharmaceuticals form a fast-growing segment in the world of medicines opening new therapeutic options for patients with severe diseases. This success is also reflected by the fast growth in global sales. Double-digit growth numbers were reported over the last 25 years, reaching $80 billion in 2012. Five drugs in the top ten of drugs with the highest sales are biopharmaceuticals (2010), clearly showing the therapeutic and economic importance of this class of drugs.

Until now biopharmaceuticals are primarily proteins, but therapeutic DNA or RNA molecules (think about gene therapy products, DNA vaccines, and RNA interference-based products; Chaps. 22, 23, and 24, respectively) may soon become part of our therapeutic arsenal.

Therapeutic proteins differ in many aspects from classical, small molecule drugs. They differ in size, composition, production, purification, contaminations, side

R.S. Oosting, Ph.D.
Division of Pharmacology, Utrecht Institute for
Pharmaceutical Sciences, Utrecht University,
Universiteitsweg 99, 3584 CG Utrecht, The Netherlands
e-mail: r.s.oosting@uu.nl

D.J.A. Crommelin, R.D. Sindelar, and B. Meibohm (eds.), *Pharmaceutical Biotechnology*,
DOI 10.1007/978-1-4614-6486-0_1, © Springer Science+Business Media New York 2013



```
a
                       signal peptide                          beta chain
human     MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
porcine   MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAEN
bovine    MALWTRLAPLLALLALWAPAPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG
          **** ** *********.* ** ******************************:***.*
                                                         alpha chain
human     LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
porcine   PQAGAVELGGGLG--GLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN
bovine    PQVGALELAGGPG-----AGGLEGPPQKRGIVEQCCASVCSLYQLENYCN
           *.* :**.** *       .***  **********:*:***********
```

Figure 1.1 ■ (**a**) Multiple alignment (http://www.ebi.ac.uk/Tools/msa/clustalw2) of the amino acid sequences of human, porcine, and bovine preproinsulin. (*): identical residue. (**b**) Schematic drawing of the structure of insulin. The alpha and beta chain are linked by two disulphide bridges. Both the one-letter and three-letter codes for the amino acids are used in this figure: alanine (*ala*, A), arginine (*arg*, R), asparagine (*asn*, N), aspartic acid (*asp*, D), cysteine (*cys*, C), glutamic acid (*glu*, E), glutamine (*gln*, Q), glycine (*gly*, H), histidine (*his*, H), isoleucine (*ile*, I), leucine (*leu*, L), lysine (*Lys*, K), methionine (*met*, M), phenylalanine (*phe*, F), proline (*pro*, P), serine (*ser*, S), threonine (*thr*, T), tryptophan (*trp*,W), tyrosine (*tyr*, Y), and valine (*val*, V) (Figure **b** is taken from Wikipedia).

effects, stability, formulation, regulatory aspects, etc. These fundamental differences justify paying attention to therapeutic proteins as a family of medicines, with many general properties different from small molecules. These general aspects are discussed in the first set of chapters of this book ("General Topics"). After those general topics, the different families of biopharmaceuticals are dealt with in detail. This first chapter should be seen as a chapter where many of the basic elements of the selection, design, and production of biopharmaceuticals are touched upon. For in detail information the reader is referred to relevant literature and other chapters in this book.

ECONOMICS AND USE

Newly introduced biopharmaceuticals are very expensive. This is partly due to the high development cost (~$1.5 billion), but this is not different from the development costs of small molecule drugs (Paul et al. 2010), combined with high production costs and, for many therapeutic proteins, a relatively low number of patients. In addition, the relatively high price of (bio) pharmaceuticals is also due to too many failures in the drug discovery and development process. The few products that actually reach the market have to compensate for all the expenses made for failed products. For a monoclonal antibody, the probability to proceed from the preclinical discovery stage into the market is around 17 % (for small molecule drugs the probability of success is even lower, ~7 %). Economic aspects of biopharmaceuticals are discussed in Chap. 10.

As mentioned above, the number of patients for many marketed therapeutic proteins is relatively small. This has several reasons. The high price of therapeutic proteins makes that they are used primarily for the treatment of the relative severe cases. The specificity of many therapeutic proteins makes that they are only effective in subgroups of patients (personalized medicine). This is in particular true for the monoclonal antibodies used to treat cancer patients. For instance, the antibody trastuzumab (Herceptin) is only approved for breast cancer patients with high expression levels of the HER2 receptor on the tumor cells (±20 % of breast cancer cases). Other examples from the cancer field are the monoclonal antibodies cetuximab and panitumumab for the treatment of metastatic colorectal cancer. Both antibodies target the EGF receptor. Successful treatment of a patient with one of these monoclonal antibodies depends on (1) the presence of the EGF receptor on the tumor and (2) the absence of mutations in signaling proteins downstream of the EGF receptor (KRAS and BRAF). Mutations in downstream signaling proteins cause the tumor to grow independently from the EGF receptor and make the tumor nonresponsive to the antagonistic monoclonal antibodies.

Some diseases are very rare and thus the number of patients is very small. Most of these rare diseases are due to a genetic defect. Examples are cystic fibrosis (CF) and glycogen storage disease II (GSD II or Pompe disease). CF is most common in Caucasians. In Europe 1:2,000–3,000 babies are affected annually. GSD II is even rarer. It affects 1:140,000 newborns. The effects of GSD II can be reduced by giving the patients recombinant myozyme. It is clear that developing a drug for such a small patient population is commercially not very interesting. To booster drug development for the rare diseases (known as orphan drugs and orphan diseases), in the USA, Europe, and Japan, specific legislation exists.

FROM AN IN SILICO DNA SEQUENCE TO A THERAPEUTIC PROTEIN

We will discuss now the steps and methods needed to select, design, and produce a recombinant therapeutic protein (see also Fig. 1.2). We will not discuss in detail the underlying biological mechanisms. We will limit ourselves, in Box 1.1, to a short description of the central dogma of molecular biology, which describes the flow of information from DNA via RNA into a protein. For detailed information, the reader is referred to more specialized molecular biology and cell biology books (see "Recommended Reading" at the end of this chapter).

■ Selection of a Therapeutic Protein

The selection of what protein should be developed for a treatment of a particular disease is often challenging, with lots of uncertainties. This is why most big phar-

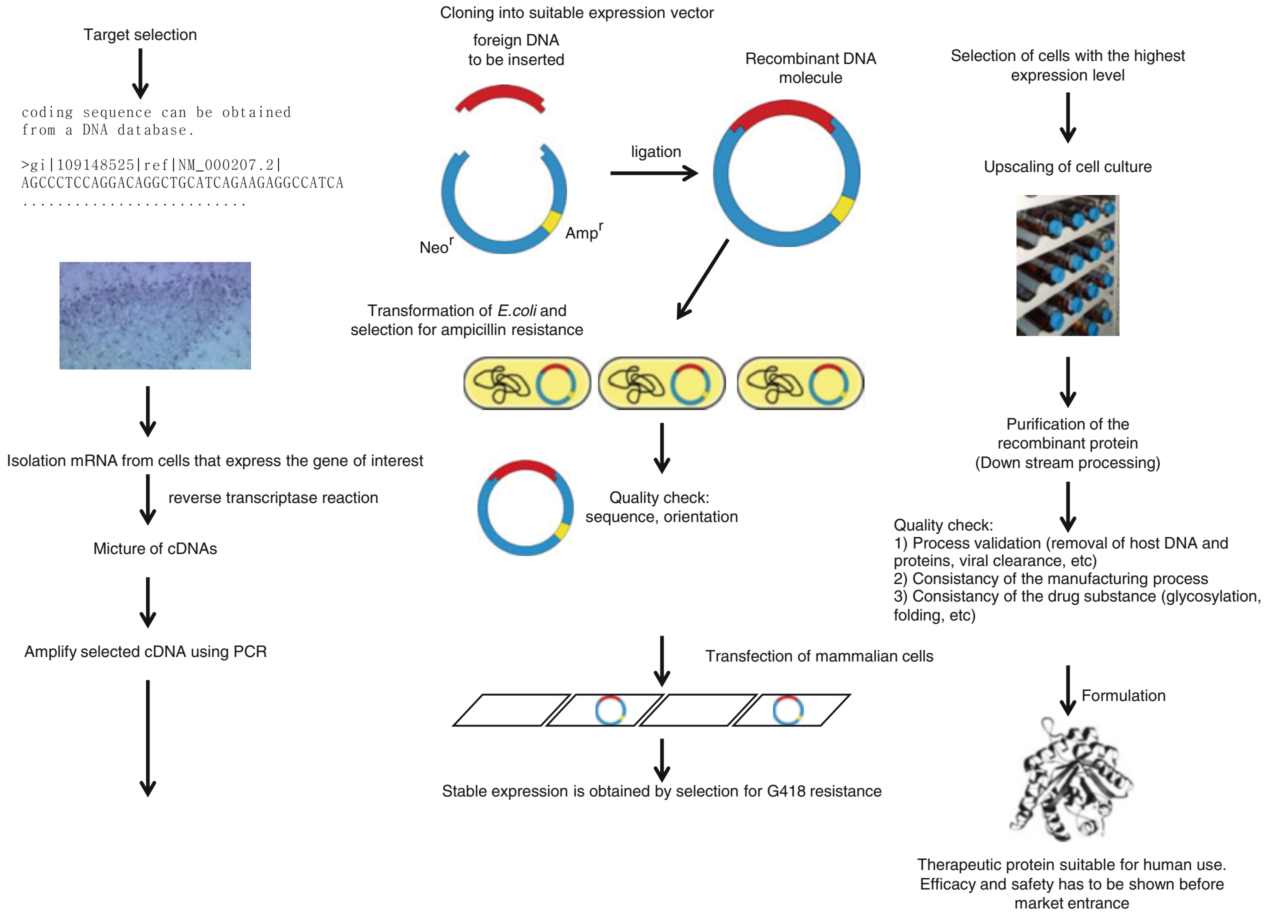
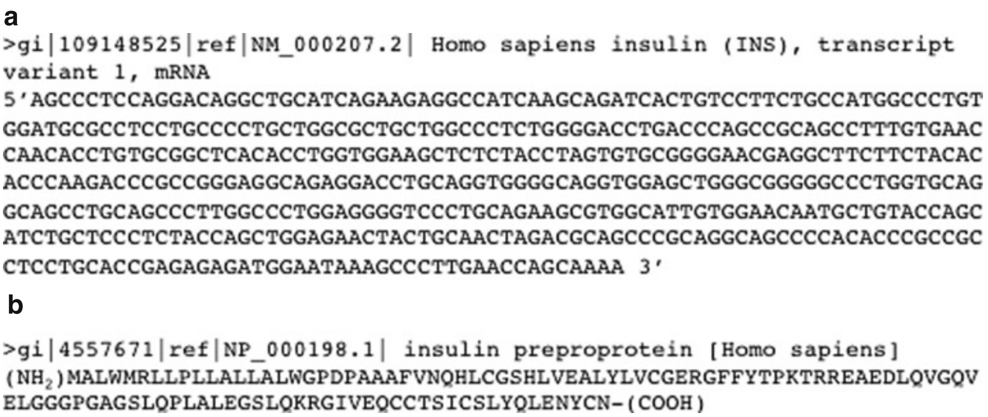

Target selection
coding sequence can be obtained
from a DNA database.
>gi|109148525|ref|NM_000207.2|
AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCA
....
Isolation mRNA from cells that express the gene of interest
reverse transcriptase reaction
Micture of cDNAs
Amplify selected cDNA using PCR
Cloning into suitable expression vector
foreign DNA
to be inserted
Neo^r
Amp^r
ligation
Recombinant DNA
molecule
Transformation of E.coli and
selection for ampicillin resistance
Quality check:
sequence, orientation
Transfection of mammalian cells
Stable expression is obtained by selection for G418 resistance
Selection of cells with the highest
expression level
Upscaling of cell culture
Purification of the
recombinant protein
(Down stream processing)
Quality check:
1) Process validation (removal of host DNA and
proteins, viral clearance, etc)
2) Consistancy of the manufacturing process
3) Consistancy of the drug substance (glycosylation,
folding, etc)
Formulation
Therapeutic protein suitable for human use.
Efficacy and safety has to be shown before
market entrance

Figure 1.2 ■ Schematic representation of all the steps required to produce a therapeutic protein.



a

```
>gi|109148525|ref|NM_000207.2| Homo sapiens insulin (INS), transcript
variant 1, mRNA
5'AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCACTGTCCTTCTGCCATGGCCCTGT
GGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAAC
CAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACAC
ACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAG
GCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGC
ATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCACACCCGCCGC
CTCCTGCACCGAGAGAGATGGAATAAAGCCCTTGAACCAGCAAAA 3'
```

b

```
>gi|4557671|ref|NP_000198.1| insulin preproprotein [Homo sapiens]
(NH2)MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQV
ELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN-(COOH)
```

Figure 1.3 ■ DNA sequences are always written from the 5′ → 3′ direction and proteins sequences from the amino-terminal to the carboxy-terminal.

maceutical companies only become interested in a certain product when there is some clinical evidence that the new product actually works and that it is safe. This business model gives opportunities for startup biotech companies and venture capitalists to engage in this important early development process.

Sometimes the choice for a certain protein as a therapeutic drug is very simple. Think, for instance, about replacement of endogenous proteins such as insulin and erythropoietin for the treatment of diabetes type I and anemia, respectively. For many other diseases it is much more difficult to identify an effective therapeutic protein or target. For instance, an antibody directed against a growth factor receptor on a tumor cell may look promising based on in vitro and animal research but may be largely ineffective in (most) human cancer patients.

It is beyond the scope of this chapter to go further into the topic of therapeutic protein and target discovery. For further information the reader is referred to the large number of scientific papers on this topic, as can be searched using PubMed (http://www.ncbi.nlm.nih.gov/pubmed).

In the rest of this chapter, we will mainly focus on a typical example of the steps in the molecular cloning process and production of a therapeutic protein. At the end of this chapter, we will shortly discuss the cloning and large-scale production of monoclonal antibodies (see also Chap. 7).

Molecular cloning is defined as the assembly of recombinant DNA molecules (most often from two different organisms) and their replication within host cells.

■ DNA Sequence

The DNA, mRNA, and amino acid sequence of every protein in the human genome can be obtained from publicly available gene and protein databases, like those present at the National Center for Biotechnology Information (NCBI) in the USA and the European Molecular Biology Laboratory (EMBL). Their websites are http://www.ncbi.nlm.nih.gov/ and http://www.ebi.ac.uk/Databases/, respectively.

DNA sequences in these databases are always given from the 5′ end to the 3′ end and protein sequences from the amino- to the carboxy-terminal end (see Fig. 1.3). These databases also contain information about the gene (e.g., exons, introns, and regulatory sequences. See Box 1.1 for explanations of these terms) and protein structure (domains, specific sites, post-translation modifications, etc.). The presence or absence of certain posttranslation modifications determines what expression hosts (e.g., *Escherichia coli* (*E. coli*), yeast, or a mammalian cell line) can be used (see below).

■ Selection of Expression Host

Recombinant proteins can be produced in *E. coli*, yeast, plants (e.g., rice and tomato), mammalian cells, and even by transgenic animals. All these expression hosts have different pros and cons.

Most marketed therapeutic proteins are produced in cultured mammalian cells. In particular Chinese hamster ovary (CHO) cells are used. On first sight, mammalian cells are not a logical choice. They are much more difficult to culture than, for instance, bacteria or yeast. On average, mammalian cells divide only once every 24 h, while cell division in *E. coli* takes ~ 30 min and in yeast ~ 1 h. In addition, mammalian cells need expensive growth media and in many cases bovine (fetal) serum as a source of growth factors (see Table 1.1 for a comparison of the various expression systems). Since the outbreak of the bovine or transmissible spongiform encephalopathy epidemic (BSE/TSE, better known as mad cow disease) under cattle in the United Kingdom, the use of bovine serum for the production of therapeutic proteins is considered a safety risk by the regulatory authorities (like the EMA in Europe and the FDA in the USA). To minimize the risk of transmitting TSE via a medicinal product, bovine serum has to be obtained from animals in countries with the lowest possible TSE risk, e.g., the USA, Australia, and New Zealand.

The main reason why mammalian cells are used as production platform for therapeutic proteins is that in these cells posttranslational modification (PTM) of the synthesized proteins resembles most closely the human situation. An important PTM is the formation of disulfide bonds between two cysteine moieties.

	Prokaryotes	**Yeast**	**Mammalian cells**
	E. coli	***Pichia pastoris Saccharomyces cerevisiae***	**(e.g., CHO or HEK293 cells)**
+	Easy manipulation Rapid growth Large-scale fermentation Simple media High yield	Grows relatively rapidly Large-scale fermentation Performs some posttranslational modifications	May grow in suspension, perform all required posttranslational modifications
–	Proteins may not fold correctly or may even aggregate (inclusion bodies) Almost no posttranslational modifications	Posttranslational modifications may differ from humans (especially glycosylation)	Slow growth Expensive media Difficult to scale up Dependence of serum (BSE)

Table 1.1 ■ Pros and cons of different expression hosts

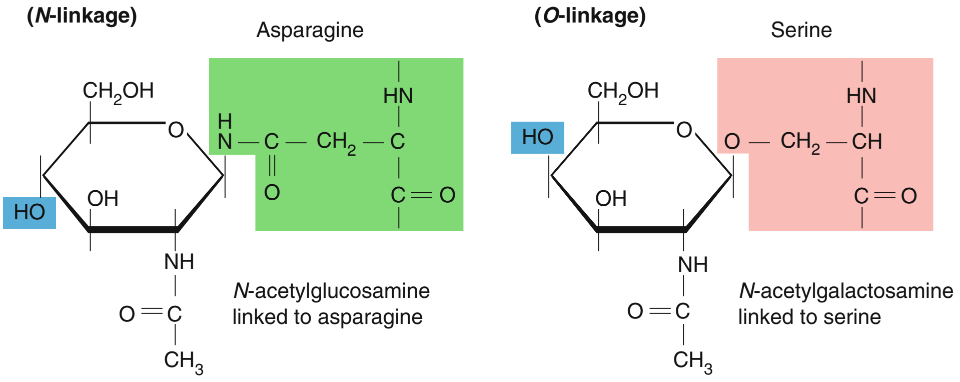


Figure 1.4 ■ Glycosylation takes place either at the nitrogen atom in the side chain of asparagine (N-linked) or at the oxygen atom in the side chain of serine or threonine. Glycosylation of asparagine takes place only when this residue is part of an Asn-X-Ser or Ans-X-Thr (X can be any residue except proline). Not all potential sites are glycosylated. Which sites become glycosylated depend also on the protein structure and on the cell type in which the protein is expressed.

Disulfide bonds are crucial for stabilizing the tertiary structure of a protein. *E. coli* is not able to make disulfide bonds in a protein, and already for this reason, *E. coli* is not very suitable for producing most of the marketed therapeutic proteins.

Another important PTM of therapeutic proteins is glycosylation. Around 70 % of all marketed therapeutic proteins, including monoclonal antibodies, are glycosylated. Glycosylation is the covalent attachment of oligosaccharides to either asparagine (N-linked) or serine/threonine (O-linked) (see Fig. 1.4). The oligosaccharide moiety of a therapeutic protein affects many of its pharmacological properties, including stability, solubility, bioavailability, in vivo activity, pharmacokinetics, and immunogenicity. Glycosylation differs between species, between different cell types within a species, and even between batches of in cell culture-produced therapeutic proteins. N-linked glycosylation is found in all eukaryotes (and also in some bacteria, but not in *E. coli*; see Nothaft and Szymanski 2010) and takes place in the lumen of the endoplasmatic reticulum and the Golgi system (see Fig. 1.5). All N-linked oligosaccharides have a common pentasaccharide core containing three mannose and two *N*-acetylglucosamine (GlcNAc) residues. Additional sugars are attached to this core. These maturation reactions take place in the Golgi system and differ between expression hosts. In yeast, the mature glycoproteins are rich in mannose, while in mammalian cells much more complex oligosaccharide structures are possible. O-linked glycosylation takes place solely in the Golgi system.

In Chap. 3 more details can be found regarding the selection of the expression system.

■ CopyDNA

The next step is to obtain the actual DNA that codes for the protein. This DNA is obtained by reverse-transcribing the mRNA sequence into copyDNA (cDNA). To explain this process, it is important to discuss first the structure of a mammalian gene and mRNA.

Most mammalian genes contain fragments of coding DNA (exons) interspersed by stretches of DNA that do not contain protein-coding information (introns). Messenger RNA synthesis starts with the making of a large primary transcript. Then, the introns are removed via a regulated process, called splicing. The mature mRNA contains only the exon sequences. Most mammalian mRNAs contain also a so-called poly-A "tail," a string of 100–300 adenosine

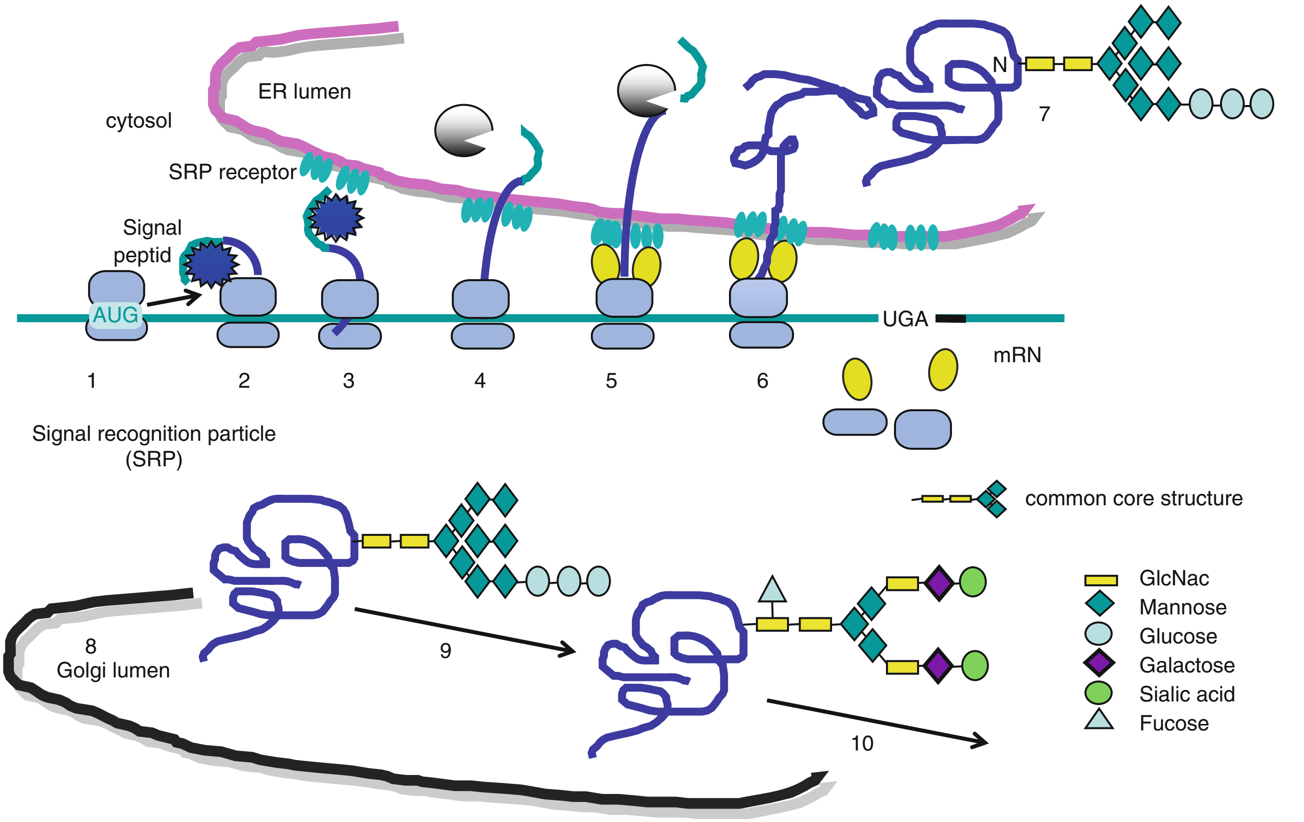


Figure 1.5 ■ Schematic drawing of the N-linked glycosylation process as occurs in the endoplasmic reticulum (ER) and Golgi system of a eukaryotic cell. (*1*) The ribozyme binds to the mRNA and translation starts at the AUG. The first ~ 20 amino acids form the signal peptide. (*2*) The signal recognition particle (SRP) binds the signal peptide. (*3*) Next, the SRP docks with the SRP receptor to the cytosolic side of the ER membrane. (*4*) The SRP is released and (*5*) the ribosomes dock onto the ER membrane. (*6*) Translation continues until the protein is complete. (*7*) A large oligosaccharide (activated by coupling to dolichol phosphate) is transferred to the specific asparagine (N) residue of the growing polypeptide chain. (*8*) Proteins in the lumen of the ER are transported to the Golgi system. (*9*) The outer carbohydrate residues are removed by glycosidases. Next, glycosyltransferases add different carbohydrates to the core structure. The complex type carbohydrate structure shown is just an example out of many possible varieties. The exact structure of the oligosaccharide attached to the peptide chain differs between cell types and even between different batches of in cell culture-produced therapeutic proteins. (*10*) Finally, secretory vesicles containing the glycoproteins are budded from the Golgi. After fusion of these vesicles with the plasma membrane, their content is released into the extracellular space.

nucleotides. These adenines are coupled to the mRNA molecule in a process called polyadenylation. Polyadenylation is initiated by binding of a specific set of proteins at the polyadenylation site at the end of the mRNA. The poly-A tail is important for transport of the mRNA from the nucleus into the cytosol, for translation, and it protects the mRNA from degradation.

An essential tool in cDNA formation is reverse transcriptase (RT). This enzyme was originally found in retroviruses. These viruses contain an RNA genome. After infecting a host cell, their RNA genome is reverse-transcribed first into DNA. The finding that RNA can be reverse-transcribed into DNA by RT is an important exception of the central dogma of molecular biology (as discussed in Box 1.1).

To obtain the coding DNA of the protein, one starts by isolating (m)RNA from cells/tissue that expresses the protein. Next, the mRNA is reverse-transcribed into copyDNA (cDNA) (see Fig. 1.6). The RT reaction is performed in the presence of an oligo-dT (a single-stranded oligonucleotide containing ~20 thymidines). The oligo-dT binds to the poly-A tail and reverse transcriptase couples deoxyribonucleotides complementary to the mRNA template, to the 3′end of the growing cDNA. In this way a so-called library of cDNAs is obtained, representing all the mRNAs expressed in the starting cells or tissue.

The next step is to amplify specifically the cDNA for the protein of interest using the polymerase chain reaction (PCR, see Fig. 1.7). A PCR reaction uses a (c) DNA template, a forward primer, a reverse primer, deoxyribonucleotides (dATP, dCTP, dGTP, and dTTP), Mg^{2+}, and a thermostable DNA polymerase. DNA polymerase adds free nucleotides only to the 3′ end of the newly forming strand. This results in elongation of the new strand in a 5′→ 3′ direction. DNA polymerase

Box 1.1 ■ The Central Dogma of Molecular Biology

The central dogma of molecular biology was first stated by Francis Crick in 1958 and deals with the information flow in biological systems and can best be summarized as "DNA makes RNA makes protein" (this quote is from Marshall Nirenberg who received the Nobel Prize in 1968 for deciphering the genetic code). The basis of the information flow from DNA via RNA into a protein is pairing of complementary bases; thus, adenine (A) forms a base pair with thymidine (*T*) in DNA or uracil in RNA and guanine (*G*) forms a base pair with cytosine (*C*).

To make a protein, the information contained in a gene is first transferred into a RNA molecule. RNA polymerases and transcription factors (these proteins bind to regulatory sequences on the DNA, like promoters and enhancers) are needed for this process. In eukaryotic cells, genes are built of exons and introns. Intron sequences (intron is derived from intragenic region) are removed from the primary transcript by a highly regulated process which is called splicing. The remaining mRNA is built solely of exon sequences and contains the coding sequence or sense sequence. In eukaryotic cells, transcription and splicing take place in the nucleus.

The next step is translation of the mRNA molecule into a protein. This process starts by binding of the mRNA to a ribosome. The mRNA is read by the ribosome as a string of adjacent 3-nucleotide-long sequences, called codons. Complexes of specific proteins (initiation and elongation factors) bring aminoacylated transfer RNAs (tRNAs) into the ribosome-mRNA complex. Each tRNAs (via its anticodon sequence) base pairs with its specific codon in the mRNA, thereby adding the correct amino acid in the sequence encoded by the gene. There are 64 possible codon sequences. Sixty-one of those encode for the 20 possible amino acids. This means that the genetic code is redundant (see Table 1.2). Translation starts at the start codon AUG, which codes for methionine and ends at one of the three possible stop codons: UAA, UGA, or UAG. The nascent polypeptide chain is then released from the ribosome as a mature protein. In some cases the new polypeptide chain requires additional processing to make a mature protein.

		2nd		**Base**			
		U	**C**	**A**	**G**		
	U	Phe	Ser	Tyr	Cys	**U**	
		Phe	Ser	Tyr	Cys	**C**	
		Leu	Ser	Stop	Stop	**A**	
1		Leu	Ser	Stop	Trp	**G**	**3**
s	**C**	Leu	Pro	His	Arg	**U**	**r**
t		Leu	Pro	His	Arg	**C**	**d**
		Leu	Pro	Gln	Arg	**A**	
b		Leu	Pro	Gln	Arg	**G**	**b**
a	**A**	Ile	Thr	Asn	Ser	**U**	**a**
s		Ile	Thr	Asn	Ser	**C**	**s**
e		Ile	Thr	Lys	Arg	**A**	**e**
		Met	Thr	Lys	Arg	**G**	
	G	Val	Ala	Asp	Gly	**U**	
		Val	Ala	Asp	Gly	**C**	
		Val	Ala	Glu	Gly	**A**	
		Val	Ala	Glu	Gly	**G**	

Table 1.2 ■ The genetic code.

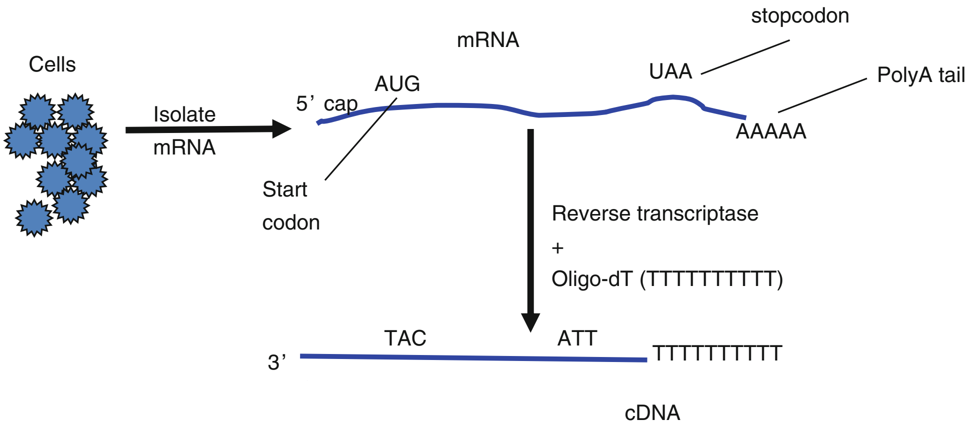


Figure 1.6 ■ Reverse transcriptase reaction.

can add a nucleotide only to a preexisting 3′-OH end, and therefore it needs a primer at which it can add the first nucleotide. PCR primers are single-stranded oligonucleotides around 20 to 30 nucleotides long, flanking opposite ends of the target DNA (see Fig. 1.8). The PCR is usually carried out for 30 cycles. Each cycle consists of three stages: a denaturing stage at ~94 °C (the double-stranded DNA is converted into single-stranded DNA), a primer annealing stage at ~60 °C (the optimal anneal temperature depends on sequences of the primers and template), and an extension stage at 72 °C. Theoretically, the amount of DNA should double during each cycle. A 30-cycle-long PCR should therefore result in a 2^{30} fold (~10^{9}) increase in the amount of DNA. In practice this is never reached. In particular at later cycles, the efficiency of the PCR reaction reduces.

PCR makes use of a thermostable DNA polymerase. These polymerases were obtained from Archaea living in hot springs such as those occurring in Yellowstone National Park (see Fig. 1.9) and at the ocean bottom. DNA polymerases make mistakes. When the aim is to clone and express a PCR product, a thermostable DNA polymerase should be used with 3′→5′ exonuclease "proofreading activity." One such enzyme is Pfu polymerase. This enzyme makes 1 mistake per every 10^{6} base pairs, while the well-known Taq polymerase, an enzyme without proofreading activity, makes on average ten times more mistakes. As a trade-off, Pfu is much slower than Taq polymerase (Pfu adds ± 1,000 nucleotides per minute to the growing DNA chain and Taq 6,000 nucleotides/min).

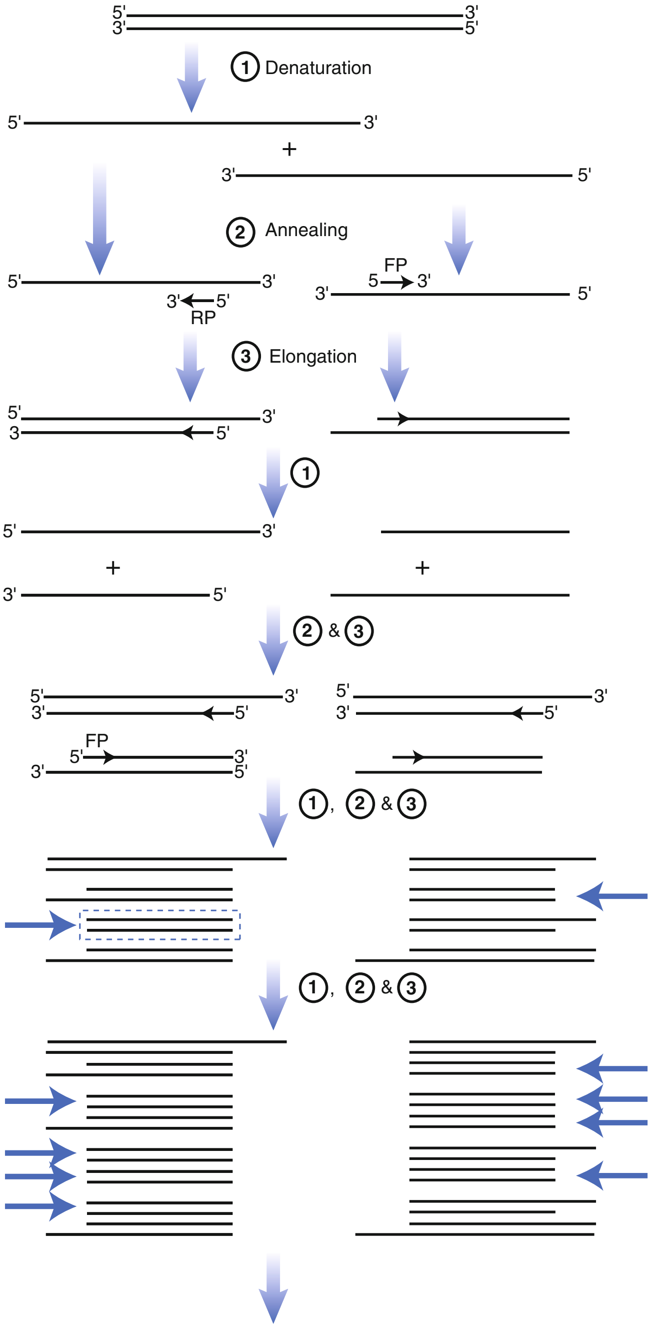


Figure 1.7 ■ The PCR process. (*1*) DNA is denatured at 94–96 °C. (*2*) The temperature is lowered to ± 60 °C. At this temperature the primers bind (anneal) to their target sequence in the DNA. (*3*) Next, the temperature is raised to 72 °C, the optimal temperature for Taq polymerase. Four cycles are shown here. A typical PCR reaction runs for 30 cycles. The *arrows* point to the desired PCR product.

■ Cloning PCR Products into an Expression Vector

There are several ways to clone a PCR product. One of the easiest ways is known as TA cloning (see Fig. 1.10). TA cloning makes use of the property of Taq polymerase to add a single adenosine to the 3′end of a PCR product. Such a PCR product can subsequently be ligated (using DNA ligase, see Molecular Biology toolbox) into a plasmid with a 5′ thymidine overhang (see Box 1.2 for a general description of expression plasmids). PCR products obtained with a DNA polymerase with proofreading activity have a blunt end, and thus they do not contain the 3′ A overhang. However, such PCR fragments can easily be A-tailed by incubating for a short period with Taq polymerase and dATP. Blunt PCR products can also directly be cloned into a linearized plasmid with 2 blunt ends. However, the efficiency of blunt-end PCR cloning is much lower than that of TA cloning. A disadvantage of TA and blunt-end cloning is that directional cloning is not possible, so the PCR fragment can be cloned either in the sense or antisense direction (see Fig. 1.10). PCR products can also be cloned by adding unique recognition sites of restriction enzymes to both ends of the PCR product. This can be done by incorporating these sites at the 5′end of the PCR primers. Although this strategy looks very straightforward, it is also not very efficient.

After ligation, the plasmid is introduced into *E. coli* by a process called transformation. There are several ways to transform *E. coli*. Most used are the calcium

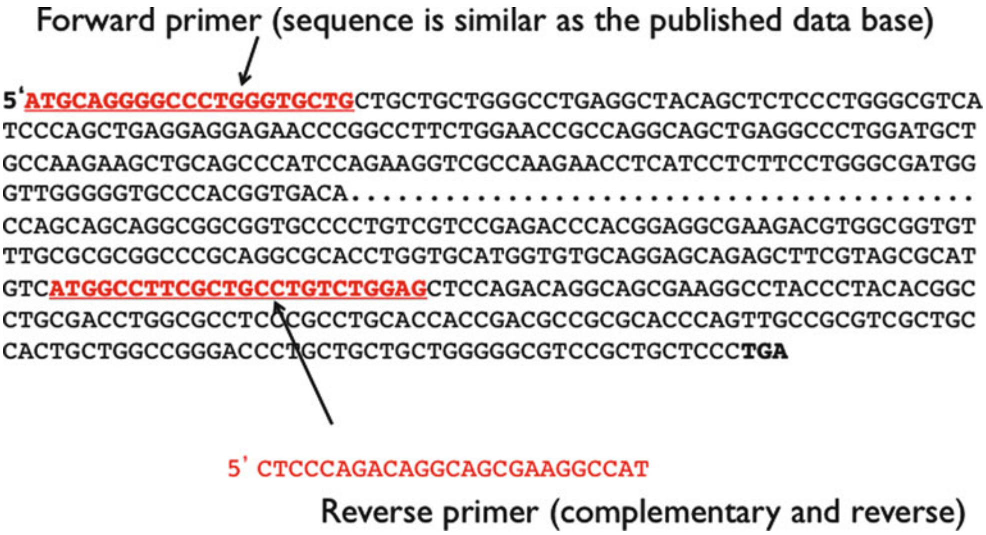

Forward primer (sequence is similar as the published data base)
5'ATGCAGGGGCCCTGGGTGCTGCTGCTGCTGGGCCTGAGGCTACAGCTCTCCCTGGGCGTCA
TCCCAGCTGAGGAGGAGAACCCGGCCTTCTGGAACCGCCAGGCAGCTGAGGCCCTGGATGCT
GCCAAGAAGCTGCAGCCCATCCAGAAGGTCGCCAAGAACCTCATCCTCTTCCTGGGCGATGG
GTTGGGGGTGCCCACGGTGACA.......................................
CCAGCAGCAGGCGGCGGTGCCCCTGTCGTCCGAGACCCACGGAGGCGAAGACGTGGCGGTGT
TTGCGCGCGGCCCGCAGGCGCACCTGGTGCATGGTGTGCAGGAGCAGAGCTTCGTAGCGCAT
GTCATGGCCTTCGCTGCCTGTCTGGAGCTCCAGACAGGCAGCGAAGGCCTACCCTACACGGC
CTGCGACCTGGCGCCTCCCGCCTGCACCACCGACGCCGCGCACCCAGTTGCCGCGTCGCTGC
CACTGCTGGCCGGGACCCTGCTGCTGCTGGGGGCGTCCGCTGCTCCCTGA
5' CTCCCAGACAGGCAGCGAAGGCCAT
Reverse primer (complementary and reverse)

Figure 1.8 ■ PCR primer design.

Figure 1.9 ■ A hot spring in Yellowstone National Park. In hot spring like this one, Archaea, the bacterial source of thermostable polymerases, live.

chloride method (better known as heat shock) and electroporation (the bacteria are exposed to a very high electric pulse). Whatever the transformation method, channels in the membrane are opened through which the plasmid can enter the cell. Next, the bacteria are plated onto an agar plate with an antibiotic. Only bacteria that have taken up the plasmid with an antibiotic-resistant gene and thus produce a protein that degrades the antibiotic will survive. After an overnight incubation at 37 °C, the agar plate will contain a number of clones. The bacteria in each colony are the descendants of one bacterium. Subsequently, aliquots of a number of these colonies are grown overnight in liquid medium at 37 °C. From these cultures, plasmids can be isolated (this is known as a miniprep). The next steps will be to determine whether the obtained plasmid preparations contain an insert, and if so, to determine what the orientation is of the insert relative to the promoter that will drive the recombinant protein expression. The orientation can, for instance, be determined by cutting the obtained plasmids with a restriction enzyme that cuts only once somewhere in the plasmid and with another enzyme that cuts once somewhere in the insert. On the basis of the obtained fragment sizes (determined via agarose gel electrophoresis using appropriate molecular weight standards), the orientation of the insert in the plasmid can be determined (see Fig. 1.10).

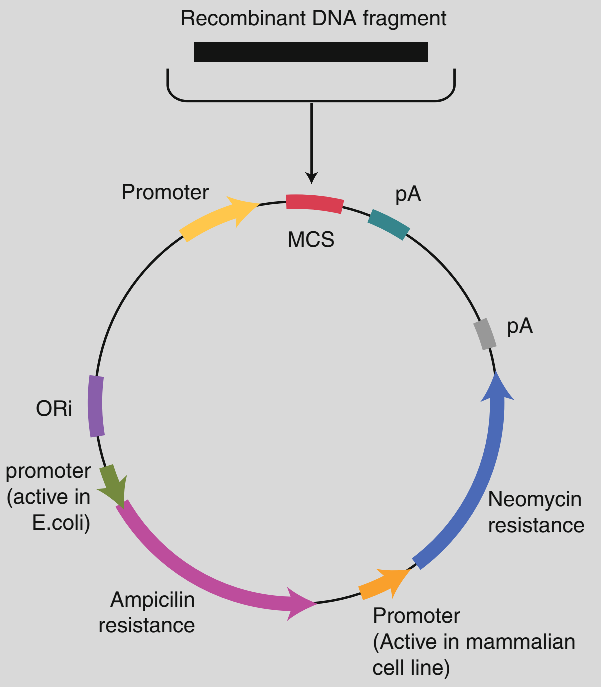
As already discussed above, DNA polymerases make mistakes, and therefore, it is crucial to determine the nucleotide sequence of the cloned PCR fragment. DNA sequencing is a very important method in biotechnology (the developments in high-throughput sequencing have enabled the sequencing of many different genomes, including that of humans) and is therefore further explained in Box 1.3.

■ Transfection of Host Cells and Recombinant Protein Production

Introducing DNA into a mammalian cell is called transfection (and as already mentioned above, transformation in *E. coli*). There are several methods to introduce DNA into a mammalian cell line. Most often, the plasmid DNA is complexed to cationic lipids (like Lipofectamine) or polymers (like polyethyleneimines or PEI) and then pipetted to the cells. Next, the positively charged aggregates bind to the negatively charged cell membrane and are subsequently endocytosized (see Fig. 1.11). Then, the plasmid DNA has to escape from the endosome and has to find its way into the nucleus where mRNA synthesis can take place. This is actually achieved during cell division when the nuclear membrane is absent. Another way to introduce DNA into the cytosol is through electroporation.

Box 1.2. ■ Plasmids.

Schematic drawing of an expression plasmid for a mammalian cell line



Plasmids are self-replicating circular extrachromosomal DNA molecules. The plasmids used nowadays in biotechnology are constructed partly from naturally occurring plasmids and partly from synthetic DNA. The figure above shows a schematic representation of a plasmid suitable for driving protein expression in a mammalian cell. The most important features of this plasmid are:

1. An origin of replication. The ori allows plasmids to replicate separately from the host cell's chromosome.
2. A multiple cloning site. The MCS contains recognition sites for a number of restriction enzymes. The presence of the MCS in plasmids makes it relatively easy to transfer a DNA fragment from one plasmid into another.
3. Antibiotic-resistant genes. All plasmids contain a gene that makes the recipient *E. coli* resistant to an antibiotic, in this case resistant to ampicillin. Other antibiotic-resistant genes that are often used confer resistance to tetracycline and Zeocin. The expression plasmid contains also the neomycin resistance gene. This selection marker enables selection of those mammalian cells that have integrated the plasmid DNA in their chromosome. The protein product of the neomycin resistance gene inactivates the toxin Geneticin.
4. Promoter to drive gene expression. Many expression vectors for mammalian cells contain the CMV promoter, which is taken from the cytomegaloma virus and is constitutively active. To drive recombinant protein expression in other expression hosts, other plasmids with other promoter sequences have to be used.
5. Poly (A) recognition site. This site becomes part of the newly produced mRNA and binds a protein complex that adds subsequently a poly (A) tail to the 3′ end of the mRNA. Expression vectors that are used to drive protein expression in *E. coli* do not contain a poly(A) recognition site.

Molecular biology enzyme toolbox

DNA polymerase produces a polynucleotide sequence against a nucleotide template strand using base pairing interactions (G against C and A against T). It adds nucleotides to a free 3′OH, and thus it acts in a 5′ → 3′ direction. Some polymerases have also 3′ → 5′ exonuclease activity (see below), which mediates proofreading.

Reverse transcriptase (RT) is a special kind of DNA polymerase, since it requires an RNA template instead of a DNA template.

Restriction enzymes are endonucleases that bind specific recognition sites on DNA and cut both strands. Restriction enzymes can either cut both DNA strands at the same location (blunt end) or they can cut at different sites on each strand, generating a single-stranded end (better known as a sticky end).

Examples:

HindIII	5′A[a]AGCTT	XhoI	5′C[a]TCGAG
	3′TTCGA[a]A		3′GAGCT[a]C
KpnI	5′GGTAC[a]C	EcoRV	5′GAT[a]ATC
	3′C^CATGG		3′CTA[a]TAG
NotI	5′GC[a]GGCCGC	PacI	5′TTAAT[a]TAA
	3′CGCCGG[a]CG		3′AAT[a]TAATT

[a] Location where the enzyme cuts

DNA ligase joints two DNA fragments. It links covalently the 3′-OH of one strand with the 5′-PO4 of the other DNA strand. The linkage of two DNA molecules with complementary sticky ends by ligase is much more efficient than blunt-end ligation.

Alkaline phosphatase. A ligation reaction of a blunt-end DNA fragment into a plasmid also with blunt ends will result primarily in empty plasmids, being the result of self-ligation. Treatment of a plasmid with blunt ends with alkaline phosphatase, which removes the 5′PO4 groups, prevents self-ligation.

Exonucleases remove nucleotides one at a time from the end (exo) of a DNA molecule. They act, depending on the type of enzyme, either in a 5′→3′ or 3→5′ direction and on single- or double-stranded DNA. Some polymerases have also exonuclease activity (required for proofreading). Exonucleases are used, for instance, to generate blunt ends on a DNA molecule with either a 3′ or 5′ extension.

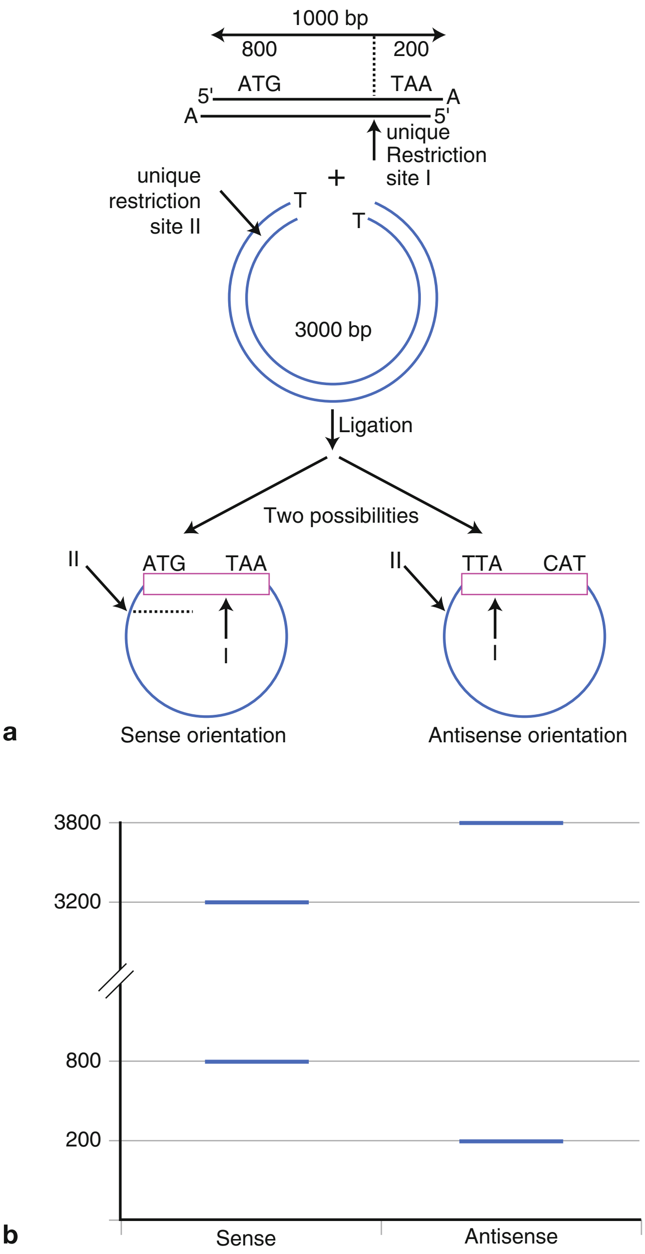


Figure 1.10 ■ Cloning of a PCR product via TA cloning (**a**). This cloning strategy makes use of the property of Taq polymerase to add an extra A to the 3′ end of the PCR product. To determine the orientation of the insert, the plasmid is cut by enzymes I and 2 (enzyme 1 cuts in the insert and enzyme 2 cuts in the plasmid). On the basis of the obtained fragment size (as determined by agarose electrophoresis), the orientation of the insert can be deduced (**b**).

During electroporation, an electric pulse is applied to the cells, which results in the formation of small pores in the plasma membrane. Through these pores the plasmid DNA can enter the cells.

Transfection leads to transient expression of the introduced gene. The introduced plasmids are rapidly diluted as a consequence of cell division or even degraded. However, it is possible to stably transfect cells leading to long expression periods. Then, the plasmid DNA has to integrate into the chromosomal DNA of the host cell. To accomplish this, a selection gene is normally included into the expression vector, which gives the transfected cells a selectable growth advantage. Only those cells that have integrated the selection

Box 1.3 ■ DNA Sequencing.

Technical breakthroughs in DNA sequencing, the determination of the nucleotide sequence, permit the sequencing of entire genomes, including the human genome. It all started with the sequencing in 1977 of the 5,386-nucleotide-long single-stranded genome of the bacteriophage φX174.

Chain-termination method and high-throughput sequencing

The most used method for DNA sequencing is the chain-termination method, also known as the dideoxynucleotide method, as developed by Frederick Sanger in the 1970s.

The method starts by creating millions of copies of the DNA to be sequenced. This can be done by isolating plasmids with the DNA inserted from bacterial cultures or by PCR. Next, the obtained double-stranded DNA molecules are denatured, and the reverse strand of one of the two original DNA strands is synthesized using DNA polymerase, a DNA primer complementary to a sequence upstream of the sequence to be determined, normal deoxynucleotidetriphosphates (dNTPs), and dideoxyNTPs (ddNTPs) that terminate DNA strand elongation. The four different ddNTPs (ddATP, ddGTP, ddCTP, or ddTTP) miss the 3’OH group required for the formation of a phosphodiester bond between two nucleotides and are each labeled with a different fluorescent dye, each emits light at different wavelengths. This reaction results in different reverse strand DNA molecules extended to different lengths. Following denaturation and removal of the free nucleotides, primers, and the enzyme, the resulting DNA molecules are separated on the basis of their molecular weight with a resolution of just one nucleotide (corresponding to the point of termination). The presence of the fluorescent label attached to the terminating ddNTPs makes a sequentially read out in the order created by the separation process possible. See also the figures below. The separation of the DNA molecules is nowadays carried out by capillary electrophoresis. The available capillary sequencing systems are able to run in parallel 96 or 384 samples with a length of 600 to 1,000 nucleotides. With the more common 96 capillary systems, it is possible to obtain around 6 million bases (Mb) of sequence per day.

Next-generation sequencing

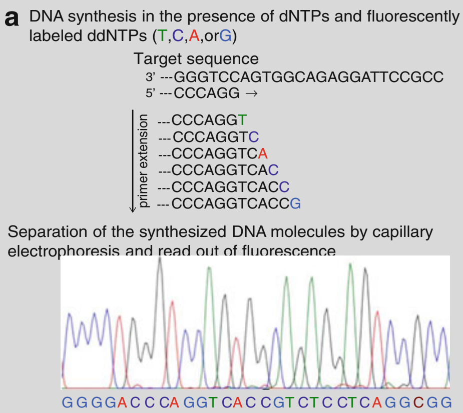
The capillary sequencing systems are still used a lot, but they will be replaced in the future by alternative systems with a much higher output (100–1,000 times more) and at the same time a strong reduction in the costs.

The description of these really high-throughput systems is beyond the purpose of this book. An excellent review about this topic is written by Kirchner and Kelso (2010).

(continued)

Box 1.3 ■ (continued)

Schematic representation of the DNA sequencing process


a
DNA synthesis in the presence of dNTPs and fluorescently labeled ddNTPs (T,C,A,orG)
Target sequence
3' ---GGGTCCAGTGGCAGAGGATTCCGCC
5' ---CCCAGG →
primer extension
---CCCAGGT
---CCCAGGTC
---CCCAGGTCA
---CCCAGGTCAC
---CCCAGGTCACC
---CCCAGGTCACCG
b
Separation of the synthesized DNA molecules by capillary electrophoresis and read out of fluorescence
G G G G A C C C A G G T C A C C G T C T C C T C A G G C G G

Schematic representation of the DNA sequencing process

marker (and most likely, but not necessary, also the gene of interest) into their genome will survive. Most expression plasmids for mammalian cells contain as selection marker the neomycin resistance gene (Neo^r). This gene codes for a protein that neutralizes the toxic drug Geneticin, also known as G418. The entire selection process takes around 2 weeks and results in a tissue culture dish with several colonies. Each colony contains the descendants of 1 stably transfected cell. Then, the cells from individual colonies have to be isolated and further expanded. The next step will be to quantify the recombinant protein production of the obtained cell cultures and to select those with the highest yields.

Transfection of mammalian cells is a very inefficient process (compared to transformation of *E. coli*) and needs relative large amounts of plasmid DNA. Integration of the transfected plasmid DNA into the genome is a very rare event. As a typical example, starting with 10^7 mammalian cells, one obtains usually not more than 10^2 stably expressing clones.

■ Cell Culture

A big challenge is to scale up cell cultures from lab scale (e.g., a 75 cm^2 tissue culture bottle) to a large-scale production platform (like a bioreactor). Mammalian cells are relatively weak and may easily become damaged by stirring or pumping liquid in or out a fermenter (shear stress). In this respect, *E. coli* is much sturdy, and thus this bacterium can therefore be grown in much larger fermenters.

A particular problem is the large-scale culturing of adherent (versus suspended) mammalian cells. One way to grow adherent cells in large amounts is on the surface of small beads. After a while the surface of the beads will be completely covered (confluent) with cells, and then, it is necessary to detach the cells from the beads and to redivide the cells over more (empty) beads and to transfer them to a bioreactor compatible with higher working volumes. To loosen the cells from the beads, usually the protease trypsin is used. It is very important that the trypsinization process is well timed: if it is too short, many cells are still on the beads, and if it is too long, the cells will lose their integrity and will not survive this.

Some companies have tackled the scale-up problem by “simply” culturing and expanding their adherent cells in increasing amounts of roller bottles. These bottles revolve slowly (between 5 and 60 revolutions per hour), which bathes the cells that are attached to the inner surface with medium (see Fig. 1.12). See Chap. 3 for more in-depth information.

■ Purification; Downstream Processing

Recombinant proteins are usually purified from cell culture supernatants or cell extracts by filtration and conventional column chromatography, including affinity chromatography (see Chap. 3).

The aim of the downstream processing (DSP) is to purify the therapeutic protein from (potential) endogenous and extraneous contaminants, like host cell proteins, DNA, and viruses.

It is important to mention here that slight changes in the purification process of a therapeutic protein may affect its activity and the amount and nature of the co-purified impurities. This is one of the main reasons (in addition to differences in expression host and culture conditions) why follow-on products (after expiration of the patent) made by a different company will never be identical to the original preparation and that is why they are not considered a true generic product (see also Chap. 11). A generic drug must contain the same active ingredient as the original drug, and in the case of a therapeutic protein, this is almost impossible and that is why the term “biosimilar” was invented.

Although not often used for the production of therapeutic proteins, recombinant protein purification may be simplified by linking it with an affinity tag, such as the his-tag (6 histidines). His-tagged proteins have a high affinity for Ni^{2+}-containing resins. There are two ways to add the 6 histidine residues. The DNA encoding the protein may be inserted into a plasmid encoding already a his-tag. Another possibility is to perform a PCR reaction with a regular primer and a primer with at its 5′end 6 histidine codons (CAT or CAC) (see Fig. 1.13). To enable easy removal of the his-tag from the recombinant protein, the tag may be followed by a suitable amino acid sequence that is recognized by an endopeptidase.

In *E. coli*, recombinant proteins are often produced as a fusion protein with another protein such as thioredoxin, beta-galactosidase, and glutathione

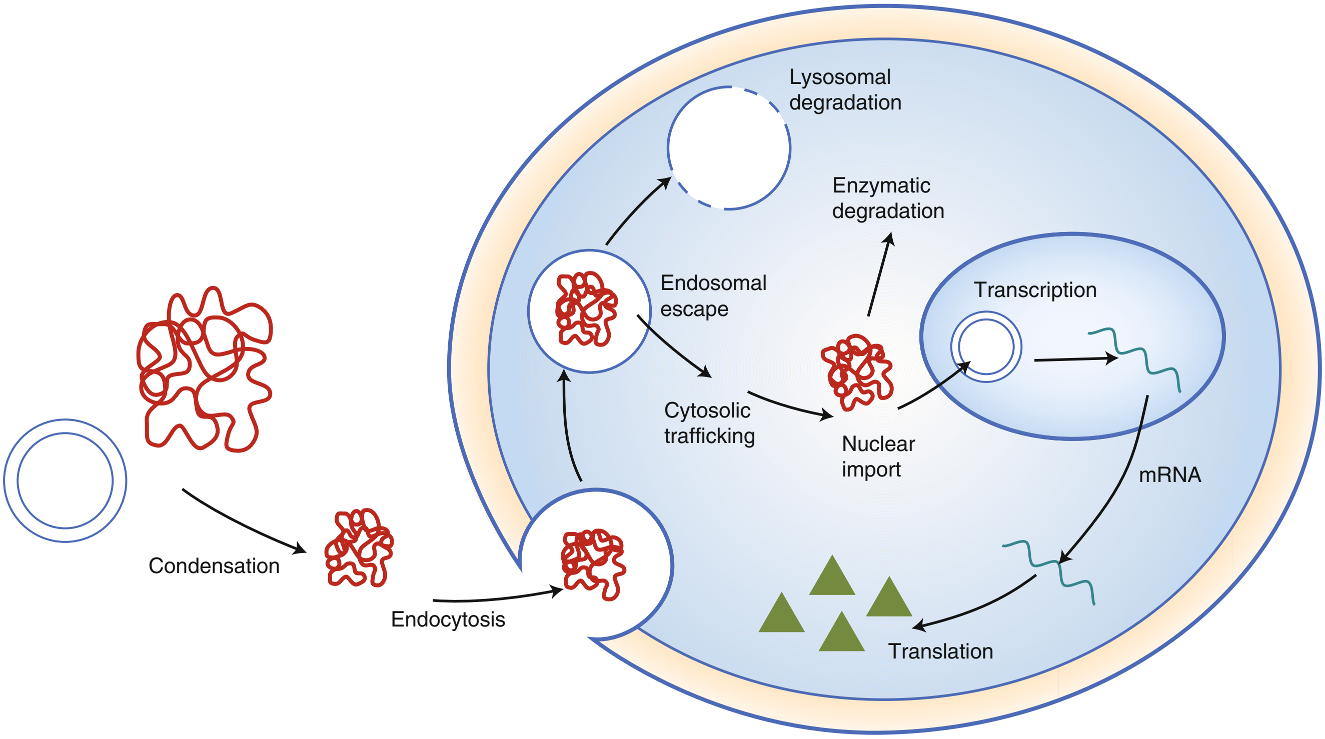


Figure 1.11 ■ Carrier-mediated transfection of mammalian cells.

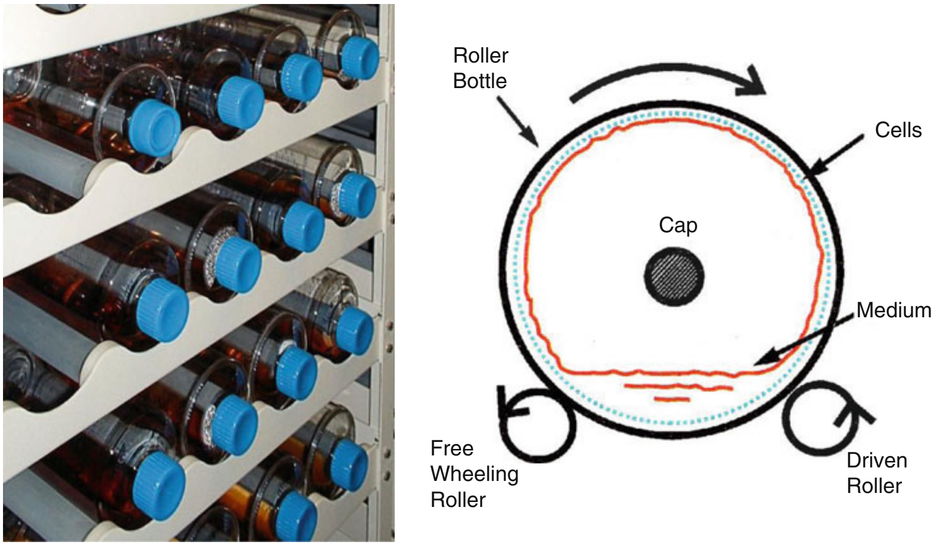


Figure 1.12 ■ Cell culturing in roller bottles.

S-transferase (GST). These fusion partners may improve the proper folding of the recombinant protein and may be used as affinity tag for purification.

MONOCLONAL ANTIBODIES

So far, we discussed the selection, design, and production of a protein starting from a DNA sequence in a genomic database. There is no database available of the entire repertoire of human antibodies. Potentially there are millions of different antibodies possible, and our knowledge about antibody-antigen interactions is not large enough to design a specific antibody from scratch.

Many marketed therapeutic proteins are monoclonal antibodies (cf. Chaps. 7, 17, 19, and 20). We will focus here on the molecular biological aspects of the design and production of (humanized) monoclonal antibodies in cell culture (primarily CHO cells are used). For a description of the structural elements of monoclonal antibodies, we refer to Chapter 7, Figs. 1.1 and 1.2.

The classic way to make a monoclonal antibody starts by immunizing a laboratory animal with a puri-

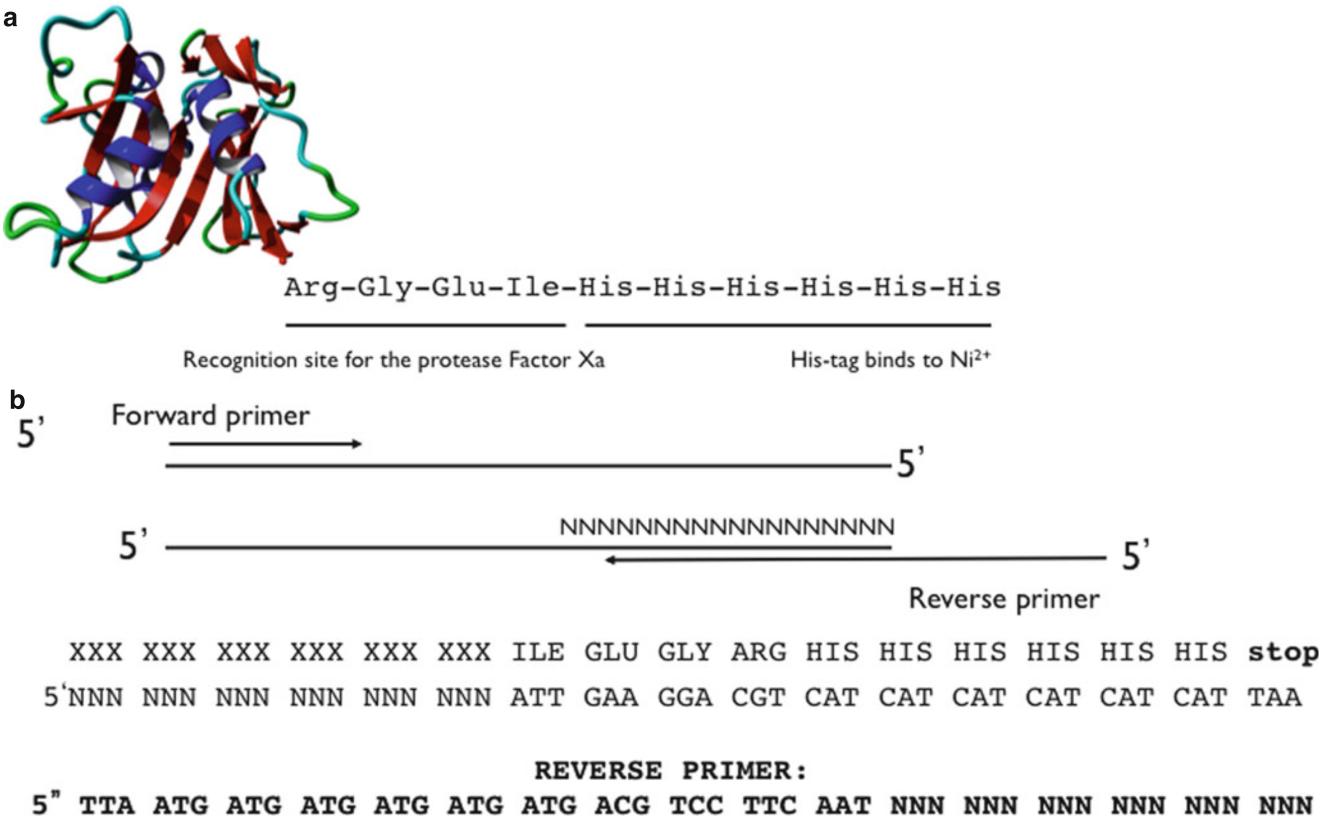

a
Arg-Gly-Glu-Ile-His-His-His-His-His-His
Recognition site for the protease Factor Xa
His-tag binds to Ni2+
b
5'
Forward primer
5'
5'
NNNNNNNNNNNNNNNNNN
5'
Reverse primer
XXX XXX XXX XXX XXX XXX ILE GLU GLY ARG HIS HIS HIS HIS HIS HIS stop
5'NNN NNN NNN NNN NNN NNN ATT GAA GGA CGT CAT CAT CAT CAT CAT CAT TAA
REVERSE PRIMER:
5" TTA ATG ATG ATG ATG ATG ATG ACG TCC TTC AAT NNN NNN NNN NNN NNN NNN

Figure 1.13 ■ (**a**) Schematic drawing of a his-tagged fusion protein. (**b**) Design of the primers needed to generate a his-tag at the carboxy-terminal end of a protein.

fied human protein against which the antibody should be directed (see Fig. 1.14). In most cases, mice are used. The immunization process (a number of injections with the antigens and an adjuvant) will take several weeks. Then the spleens of these mice are removed and lymphocytes are isolated. Subsequently, the lymphocytes are fused using polyethylene glycol (PEG) with a myeloma cell. The resulting hybridoma cell inherited from the lymphocytes the ability to produce antibodies and from the myeloma cell line the ability to divide indefinitely. To select hybridoma cells from the excess of non-fused lymphocytes and myeloma cells, the cells are grown in HAT selection medium. This culture medium contains hypoxanthine, aminopterin, and thymidine. The myeloma cell lines used for the production of monoclonal antibodies contain an inactive hypoxanthine-guanine phosphoribosyltransferase (HGPRT), an enzyme necessary for the salvage synthesis of nucleic acids. The lack of HGPRT activity is not a problem for the myeloma cells because they can still synthesize purines de novo. By exposing the myeloma cells to the drug aminopterin also de novo synthesis of purines is blocked and these cells will not survive anymore. Selection against the unfused lymphocytes is not necessary, since these cells, like most primary cells, do not survive for a long time in cell culture. After PEG treatment, the cells are diluted and divided over several dishes. After approximately 2 weeks, individual clones are visible. Each clone contains the descendants of one hybridoma cell and will produce one particular type of antibody (that is why they are called monoclonal antibodies). The next step is to isolate hybridoma cells from individual clones and grow them in separate wells of a 96-well plate. The hybridomas secrete antibodies into the culture medium. Using a suitable test (e.g., an ELISA), the obtained culture media can be screened for antibody binding to the antigen. The obtained antibodies can then be further characterized using other tests. In this way a mouse monoclonal antibody is generated.

These mouse monoclonal antibodies cannot be used directly for the treatment of human patients. The amino acid sequence of a mouse antibody is too different from the sequence of an antibody in humans and thus will elicit an immune response. To make a mouse antibody less immunogenic, the main part of its sequence must be replaced by the corresponding human sequence. Initially, human-mouse chimeric antibodies were made. These antibodies consisted of the constant regions of the human heavy and light chain and the variable regions of the mouse antibody.

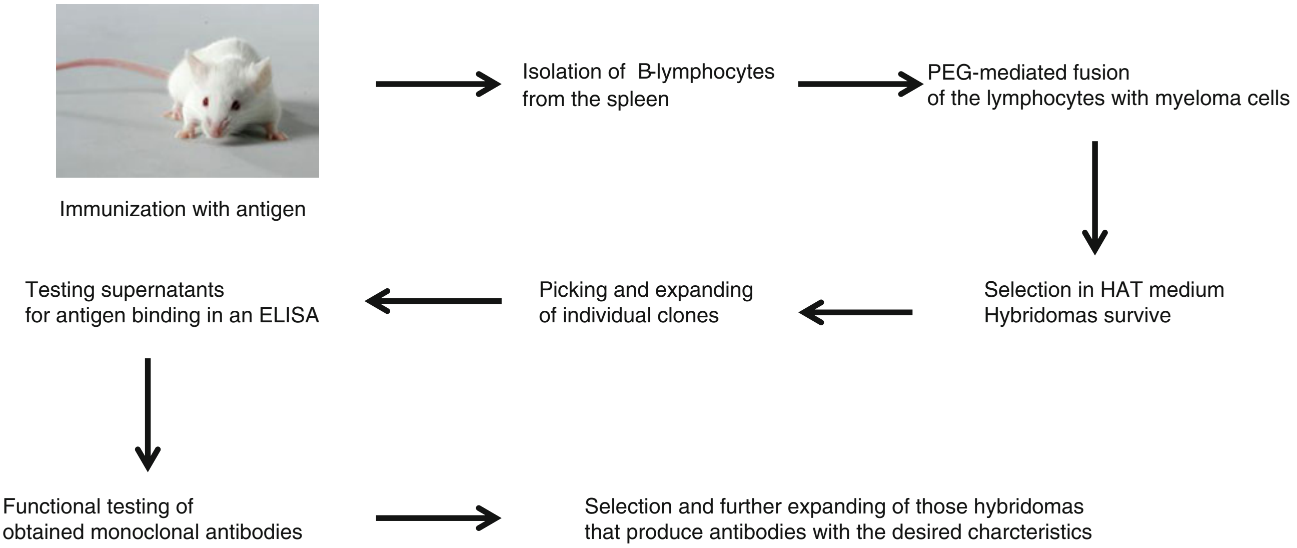

Immunization with antigen
Isolation of B-lymphocytes from the spleen
PEG-mediated fusion of the lymphocytes with myeloma cells
Selection in HAT medium Hybridomas survive
Picking and expanding of individual clones
Testing supernatants for antigen binding in an ELISA
Functional testing of obtained monoclonal antibodies
Selection and further expanding of those hybridomas that produce antibodies with the desired charcteristics

Figure 1.14 ■ The making of a mouse monoclonal antibody.

Later, so-called humanized antibodies were generated by grafting only the complementarity-determining regions (CDRs), which are responsible for the antigen-binding properties, of the selected mouse antibody onto a human framework of the variable light (V_L) and heavy (V_H) domains. The humanized antibodies are much less immunogenic than the previously used chimeric antibodies. To even further reduce immunogenicity, SDR grafting is used nowadays (Kashmiri et al. 2005). SDR stands for 'specificity determining residues'. From the analysis of the 3-D structure of antibodies, it appeared that only ~30 % of the amino acid residues present in the CDRs are critical for antigen binding. These residues, which form the SDR, are thought to be unique for a given antibody.

Humanization of a mouse antibody is a difficult and tricky process. It results usually in a reduction of the affinity of the antibody for its antigen. One of the challenges is the selection of the most appropriate human antibody framework. This framework determines basically the structure of the antibody and thus the orientation of the antigen recognition domains in space. Sometimes it is necessary to change some of the residues in the human antibody framework to restore antigen binding. To further enhance the affinity of the humanized antibody for its antigen, mutations within the CDR/SDR sequences are introduced. How this in vitro affinity maturation is done is beyond the scope of this chapter.

So far the generation of a humanized antibody has been described in a rather abstract way. How is it done in practice? First, the nucleotide sequence of each of the V_L and V_H regions is deduced (contains either the murine CDRs or SDRs). Next, the entire sequence is divided over four or more alternating oligonucleotides with overlapping flanks (see Fig. 1.15). These relatively long oligonucleotides are made synthetically. The reason why the entire sequence is divided over four nucleotides instead of over two or even one is that there is a limitation to the length of an oligonucleotide that can be synthesized reliably (a less than 100 % yield of each coupling step (nucleotides are added one at the time) and the occurrence of side reactions make that oligonucleotides hardly exceed 200 nucleotide residues).

To the four oligonucleotides, a heat-stable DNA polymerase and the 4 deoxyribonucleotides (dATP, dCTP, dGTP, and dTTP) are added, and the mixture is incubated at an appropriate temperature. Finally 2 primers, complementary to both ends of the fragment, are added, which enable the amplification of the entire sequence. The strategy to fuse overlapping oligonucleotides by PCR is called PCR sewing.

Finally, the PCR product encoding the humanized V_L and V_H region is cloned into an expression vectors carrying the respective constant regions and a signal peptide. The signal peptide is required for glycosylation. Subsequently, the expression constructs will be used to stably transfect CHO cells. The obtained clones will be tested for antibody production, and clones with the highest antibody production capacity will be selected for further use.

YIELDS

To give an idea about the production capacity needed to produce a monoclonal antibody, we will do now some calculations. The annual production of the most successful therapeutic monoclonal antibodies is around 1,000 kg (in 2009). In cell culture, titers of 2–6 g/L are routinely reached and the yield of the DSP is around 80 %. Thus, to produce 1,000 kg monoclonal antibody,

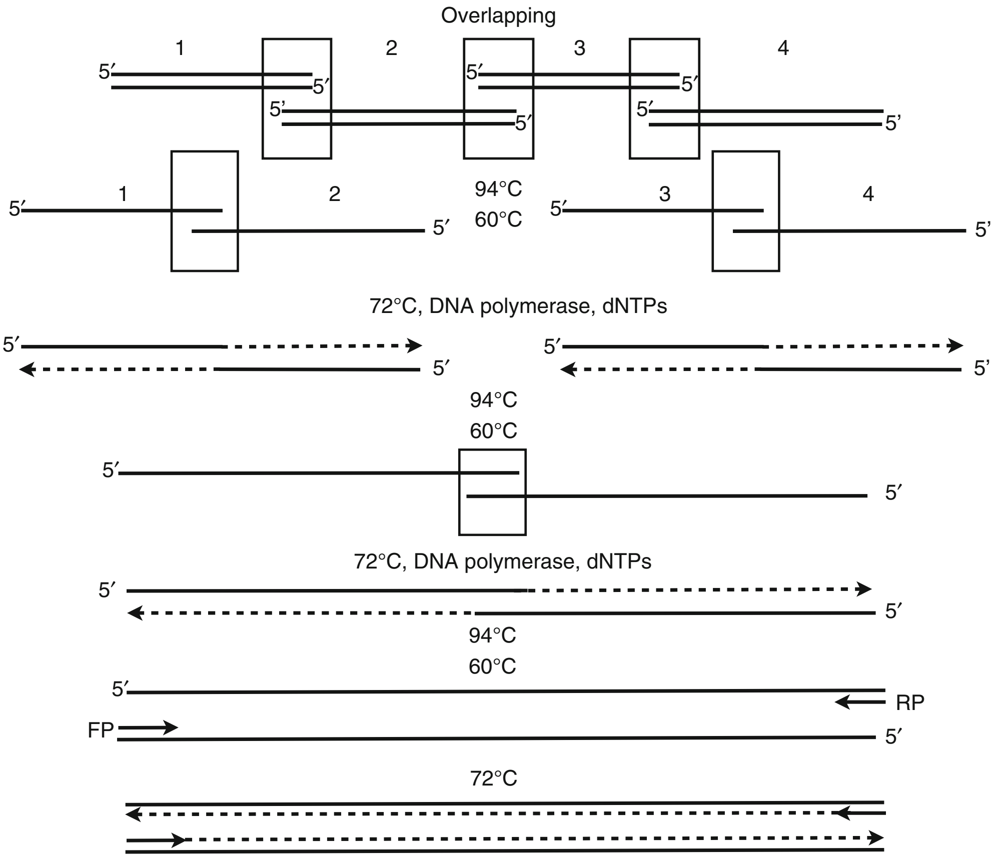


Fig. 1.15 ■ The making of a sequence containing the humanized V_L or V_H region of an antibody by PCR sewing. Both the V_L and V_H regions contain three highly variable loops (known as complementarity-determining regions 1, 2, and 3). The V_L and V_H regions are approximately 110 amino acids in size. Four alternating oligonucleotides with overlapping flanks are incubated together with a DNA polymerase and deoxynucleotides. DNA polymerase fills in the gaps. The sequences of these oligonucleotides are based upon the original mouse CDR/SDR sequences inserted into a human V_L or V_H framework. Next, the entire sequence is PCR amplified using end primers (*FP* forward primer, *RP* reverse primer). The resulting PCR fragments will be around 330 base pair in size

one needs 200,000–600,000 L of cell culture supernatant. In Chap. 3 more details can be found.

CONCLUSION

Thanks to advances in many different areas, including molecular biology, bioinformatics, and bioprocess engineering, we have moved from an animal-/human-derived therapeutic protein product towards in vitro-produced therapeutic proteins with the fully human sequence and structure. Importantly, we have now access to potentially unlimited amounts of high-quality therapeutic proteins. Of course, there will always be a risk for (viral) contaminations in the *in vitro*-produced therapeutic protein preparation, but this risk is much smaller than when the protein has to be isolated from a human source (examples from the past include the transmission of hepatitis B and C and HIV via blood-derived products and the transmission of Creutzfeldt-Jakob disease from human growth hormone preparations from human pituitaries).

As basic knowledge in molecular biology and engineering keeps on growing, the efficiency of the cloning and production process will increase in parallel.

SELF-ASSESSMENT QUESTIONS

■ Questions

1. A researcher wanted to clone and subsequently express the human histone H4 protein in *E. coli*.

   She obtained the sequence below from the NCBI, as shown below. The start and stop codons are underlined.

   >gi|29553982|ref|NM_003548.2| Homo sapiens histone cluster 2, H4a (HIST2H4A), mRNA

AGAAGCTGTCTATCGGGCTCCAGCGGTCATGTCCG
GCAGAGGAAAGGGCGGAAAAGGCTTAGGCAA
AGGG
GGCGCTAAGCGCCACCGCAAGGTCTTGAGAGAC
AACATTCAGGGCATCACCAAGCCTGCCATTCG
GCGTC

TAGCTCGGCGTGGCGGCGTTAAGCGGATCTCTGGCCTCATTTACGAGGAGACCCGCGGTGTGCTGAAGGT
GTTCCTGGAGAATGTGATTCGGGACGCAGTCACCTACACCGAGCACGCCAAGCGCAAGACCGTCACAGCC
ATGGATGTGGTGTACGCGCTCAAGCGCCAGGGGCGCACCCTGTACGGCTTCGGAGGCTAGGCCGCCGCTC
CAGCTTTGCACGTTTCGATCCCAAAGGCCCTTTTTAGGGCCGACCA.

(i) Is *E. coli* a suitable expression host for the H4 protein?

(ii) Design primers for the amplification of the coding sequence of the H4 protein by PCR.

(iii) To ease purification the researcher decided to add an affinity tag (Trp-Ser-His-Pro-Gln-Phe-Glu-Lys) to the carboxy-terminal end of the H4 protein. PCR was used to clone this tag in frame with the H4 protein. What was the sequence of the primers she probably used?
To answer this question, make use of the table below.

		2nd		Base			
		U	C	A	G		
	U	Phe	Ser	Tyr	Cys	U	
		Phe	Ser	Tyr	Cys	C	
		Leu	Ser	Stop	Stop	A	
1		Leu	Ser	Stop	Trp	G	3
s	C	Leu	Pro	His	Arg	U	r
t		Leu	Pro	His	Arg	C	d
		Leu	Pro	Gln	Arg	A	
b		Leu	Pro	Gln	Arg	G	b
a	A	Ile	Thr	Asn	Ser	U	a
s		Ile	Thr	Asn	Ser	C	s
e		Ile	Thr	Lys	Arg	A	e
		Met	Thr	Lys	Arg	G	
	G	Val	Ala	Asp	Gly	U	
		Val	Ala	Asp	Gly	C	
		Val	Ala	Glu	Gly	A	
		Val	Ala	Glu	Gly	G	

(iv) And finally, she decided to optimize the codon usage for expression in *E. coli*. What is coding optimalization? What is its purpose?

(v) Design a strategy/method to optimize the codon usage of the H4 protein.

(vi) The human H4 mRNA differs from most other human mRNAs by lacking a poly-A tail (instead the H4 mRNA is protected by a palindromic termination element), and thus the cDNA encoding this protein cannot be obtained by a reverse transcriptase reaction using an oligo-dT as primer. Describe a method to obtain the H4 cDNA.

2. Ampicillin, G418, and HAT medium are used to select for transformed *E. coli*, transfected mammalian cells, and hybridomas, respectively. Describe shortly the mechanism underlying the three mentioned selection strategies.

3. *E. coli* does not take up plasmid DNA spontaneously. However, the so-called chemical competent *E. coli* is able to take up plasmids following a heat shock (30 s 42 °C, followed by an immediate transfer to 0 °C). These competent bacteria can be obtained by extensive washing with a 100 mM $CaCl_2$ solution.

Transformation of competent *E. coli* of good quality results in ± 10^8 colonies/μg of supercoiled plasmid DNA. The bacteria in each colony are the descendants of one bacterium that had initially taken up one plasmid molecule.

Calculate the transformation efficiency defined as the number of plasmids taken up by the competent bacteria divided by the total number of plasmids added. Make the calculation for a plasmid of 3,333 base pairs (the MW of a nucleotide is 300 g/mol and the Avogadro constant is 6×10^{23} molecules/mol).

■ Answers

1. (i) Information about the protein structure can be obtained from http://www.expasy.org/. The H4 protein does not contain disulfide bridges and is unglycosylated. It is therefore likely that *E. coli* is able to produce a correctly folded H4 protein.

(ii) PCR primers are usually around 18–20 nucleotides long. The sequences of the forward and reverse primer are *ATG* TCC GGC AGA GGA AAG (identical to the published sequence) and *CTA* GCC TCC GAA GCC GTA (complementary and reverse), respectively.

(iii) The forward primer will be as above. At the 5′ end of the reverse primer, additional sequences must be added. First, the DNA sequence encoding the affinity tag Trp-Ser-His-Pro-Gln-Phe-Glu-Lys must be determined using the codon usage table: TCG CAC CCA CAG TTC GAA AAG. It is important to place the tag in front of the stop codon (TAG). The sequence of the reverse primer will then be 5′- CTA CTT TTC GAA CTG TGG GTG CGA CCA GCC TCC GAA GCC GTA CAG- 3′.

(iv) For most amino acids more than one codon exist (see the codon usage table).
Differences in preferences for one of the several codons that encode the same amino acid exist between organisms. In particular in fast-growing organisms, like *E. coli*, the optimal codons reflect the composition of their transfer RNA (tRNA) pool. By changing the native codons into those codons preferred by *E. coli*, the level of heterologous protein expression may increase. Alternatively, and much easier, one could use as expression host an *E. coli* with plasmids encoding extra copies of rare tRNAs.

(v) The H4 protein is 103 amino acids long. The easiest way to change the sequence at many places along the entire length of the coding sequence/mRNA is by designing four overlapping oligonucleotides.

Next, the four overlapping oligonucleotides must be "sewed" together by a DNA polymerase in the presence of dNTPs. Finally, by the addition of two flanking primers, the entire, now optimized sequence can be amplified.

(vi) An oligo-dT will not bind to the mRNA of H4, and therefore one has to use a H4-specific primer. One could use for instance the reverse primer as designed by question 1.ii

2. (a) *Selection of transformed bacteria using ampicillin*. The antibiotic ampicillin is an inhibitor of transpeptidase. This enzyme is required for the making of the bacterial cell wall. The ampicillin resistance gene encodes for the enzyme beta-lactamase, which degrades ampicillin.

(b) *Selection of stably transfected mammalian cells using G418*. Most expression plasmids for mammalian cells contain as selection marker the neomycin resistance gene (Neo^r). This gene codes for a protein that neutralizes the toxic drug Geneticin, also known as G418. G418 blocks protein synthesis both in prokaryotic and eukaryotic cells. Only cells that have incorporated the plasmid with the Neo^r gene into their chromosomal DNA will survive.

(c) *Selection of hybridomas using HAT medium*. HAT medium contains hypoxanthine, aminopterin, and thymidine. The myeloma cell lines used for the production of monoclonal antibodies contain an inactive hypoxanthine-guanine phosphoribosyltransferase (HGPRT), an enzyme necessary for the salvage synthesis of nucleic acids. The lack of HGPRT activity is not a problem for the myeloma cells because they can still synthesize purines de novo. By exposing the myeloma cells to the drug aminopterin also de novo synthesis of purines is blocked and these cells will not survive anymore. Selection against the unfused lymphocytes is not necessary, since these cells, like most primary cells, do not survive for a long time in cell culture.

3. First, calculate the molecular weight of the plasmid: $333 \times 2 \times 300 = 2 \times 10^6$ g/mol. → 2×10^6 g plasmid $= 6 \times 10^{23}$ molecules. → 1 g plasmid $= 3 \times 10^{17}$ molecules. →1 µg (1×10^{-6} g) $= 3 \times 10^{11}$ molecules.

1 µg gram plasmid results in 10^8 colonies. Thus, only one in 3,000 plasmids is taken up by the bacteria.

RECOMMENDED READING AND REFERENCES

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2007) Molecular biology of the cell. Garland Science, New York

Berg JM, Tymoczko JL, Stryer L (2011) Biochemistry, 7th edn. WH. Freeman & CO., New York

Brekke OH, Sandlie I (2003) Therapeutic antibodies for human diseases at the dawn of the twenty-first century. Nat Rev Drug Discov 2(1):52–62

Wikepedia. Available at: http://en.wikipedia.org

Kashmiri SV, De Pascalis R, Gonzales NR, Schlom J (2005) SDR grafting-a new approach to antibody humanization. Methods 36(1):25–34

Kircher M, Kelso J (2010) High-throughput DNA sequencing–concepts and limitations. Bioessays 32(6):524–536

Leader B, Baca QJ, Golan DE (2008) Protein therapeutics: a summary and pharmacological classification. Nat Rev Drug Discov 7(1):21–39

Lodish H, Berk A, Kaiser CA, Krieger M, Scott MP (2007) Molecular cell biology, 6th edn. WH. Freeman & CO., New York

Nothaft H, Szymanski CM (2010) Protein glycosylation in bacteria: sweeter than ever. Nat Rev Microbiol 8(11):765–778

Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9(3):203–214

Strohl WR, Knight DM (2009) Discovery and development of biopharmaceuticals: current issues. Curr Opin Biotechnol 20(6):668–672