Protein Structures and Structure-Based Rational Drug Design

P.M. Selzer et al., *Applied Bioinformatics*, https://doi.org/10.1007/978-3-319-68301-0_5

5.1 Protein Structure

Proteins are macromolecules whose monomeric subunits are the naturally occurring 20 amino acids. The amino acids are linked via peptide bonds (generated upon water release) to form a polypeptide (▶ Chap. 1). Polypeptides can range in length from three to several hundred amino acids. The amino acid sequence of a given protein, also known as the primary structure, is genetically determined. It becomes fixed during translation based on the information encoded in the mRNA.

The properties of an extended polypeptide chain correspond to a cross section of those of the corresponding amino acids, i.e., the function of the corresponding protein cannot be determined solely from the primary structure. It also depends on the spatial arrangement of the amino acids to one another. Stretched polypeptide chains fold spontaneously into secondary structures and then into three-dimensional (3D) structures. The secondary structure can comprise two main structural features, the α-helix and the β-sheet. These structural elements are connected via nonrepetitive elements called loops, which consist of irregular turns as third secondary structural elements. It is the combination of the positioning of the amino acid side chains and the peptide backbone of the secondary structure that forms the protein tertiary structure. If a protein consists of several subunits, then the association of these subunits to form the functional protein is called the quaternary structure.

The function of a protein is mediated by its 3D structure, which, if known, can allow the inference of function. A reliable *ab initio* prediction of protein tertiary structure based solely on the primary structure is not yet possible, at least in the near future. Also, the experimental determination of structure is still difficult and the number of known protein structures comparatively small. Therefore, the prediction of the protein function based on the tertiary or quaternary structure is limited. However, proteins show a variety of structural and topological features that can be used to predict their properties and functions. Many of these features can be inferred or predicted from the primary structure by computational methods. Some of these properties and their predictions are discussed in what follows.

5.2 Signal Peptides

For many proteins the site of synthesis is not the site of action. This applies to transmembrane proteins, proteins within the endoplasmic reticulum, and proteins that are secreted or imported into lysosomes. Prior to their activation, these proteins must first be transported to the site of action, and this is facilitated by a peptide recognition signal for the cellular transport system. The recognition signal is an N-terminal leader sequence (signal peptide) that consists of approx. 15–30 amino acids placed on the N-terminus of the mature protein (▣ Fig. 5.1). According to the signal hypothesis of Günter Blobel and David Sabatini (Blobel and Sabatini 1971), the signal peptide is recognized by a signal recognition particle, guiding the nascent polypeptide chain through the membrane of the endoplasmic reticulum. As soon as the signal peptide has passed the membrane, it is specifically cleaved from the nascent polypeptide by a signal peptidase. Proteins with

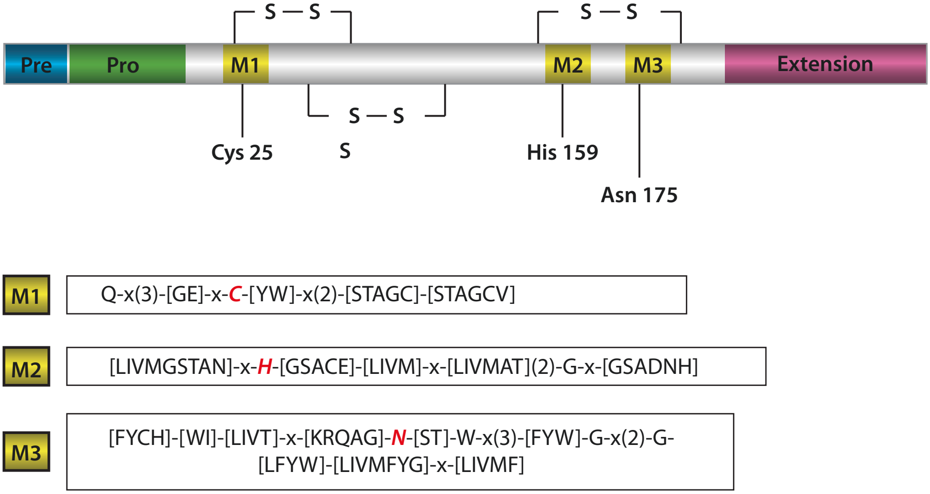

S — S
S — S
Pre
Pro
M1
M2
M3
Extension
S — S
Cys 25
S
His 159
Asn 175
M1
Q-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC]-[STAGCV]
M2
[LIVMGSTAN]-x-H-[GSACE]-[LIVM]-x-[LIVMAT](2)-G-x-[GSADNH]
M3
[FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-
[LFYW]-[LIVMFYG]-x-[LIVMF]

Fig. 5.1 Schematic illustration of a preproprotein exemplified by cysteine proteases of the papain family. The amino acids of the catalytic triad (Cys25, His159, and Asp175) are each located within the characteristic sequence motifs of cysteine proteases (M1–M3). Only a few cysteine proteases have an additional C-terminal extension for which a function is still not known

a signal peptide are called preproteins or, in those cases where they also contain propeptides, preproproteins. Unlike signal peptides, propeptides are proteolytically removed to allow for protein activation (Fig. 5.1).

The presence of a signal peptide gives an important clue as to the site of action of proteins. This knowledge in turn can help clarify function and, thus, help in determining whether that protein is a suitable target molecule. For this reason, methods for predicting the presence of signal peptides in the primary structure have been developed. An example is the program SignalP from the Center for Biological Sequence Analysis (CBS) at the Technical University of Denmark [signalp] (Petersen et al. 2011). The recognition of signal peptides by the signal recognition particle is not due to a conserved amino acid sequence but depends on physicochemical properties. A signal peptide usually consists of three parts. The first region (the n-region) contains 1–5 usually positively charged amino acids, the second region (the h-region) is made up of 5–15 hydrophobic amino acids, and the third region (the c-region) has 3–7 polar but mostly uncharged amino acids. A classical sequence alignment method is therefore unsuitable for the prediction of signal peptides. The SignalP program in its current fourth version is instead based on the use of neural networks. Using machine learning methods, the characteristics of a training data set with known sequences are learned and can be used for the prediction of unknown data. The trained neural networks are thus able to judge the properties of amino acids in unknown sequences, thereby allowing the recognition of signal sequences. SignalP uses two different neural networks since signal peptides and transmembrane helices (▶ Sect. 5.5.3) can hardly be differentiated. One neural network is therefore trained with signal peptide sequences, while

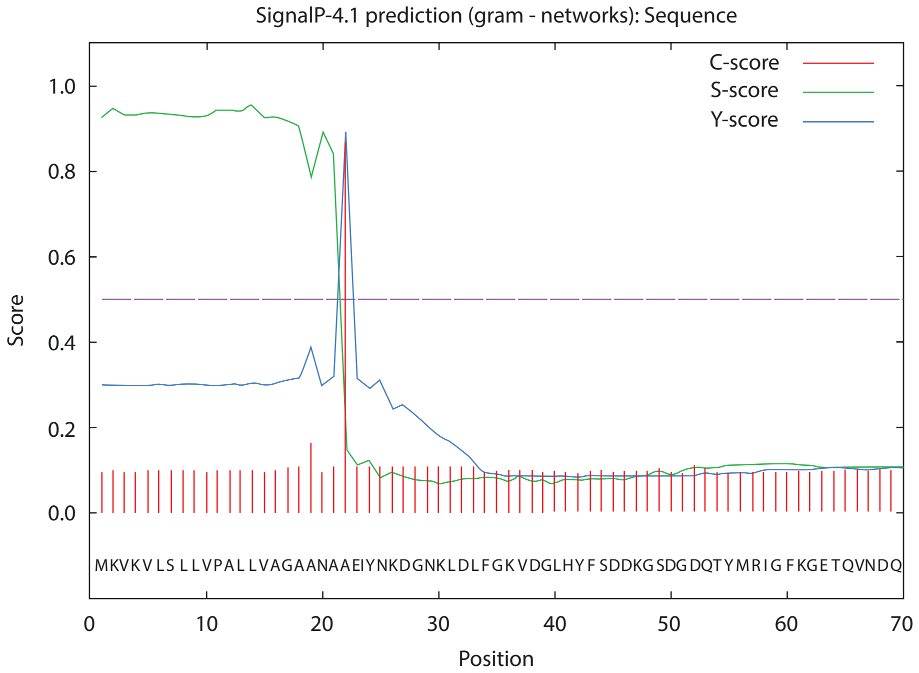


Fig. 5.2 Graphical output of SignalP server [signalp] at CBS

the other is trained with sequences of transmembrane helices. Using this approach, the false positive rate of predicted signal peptides could be minimized.

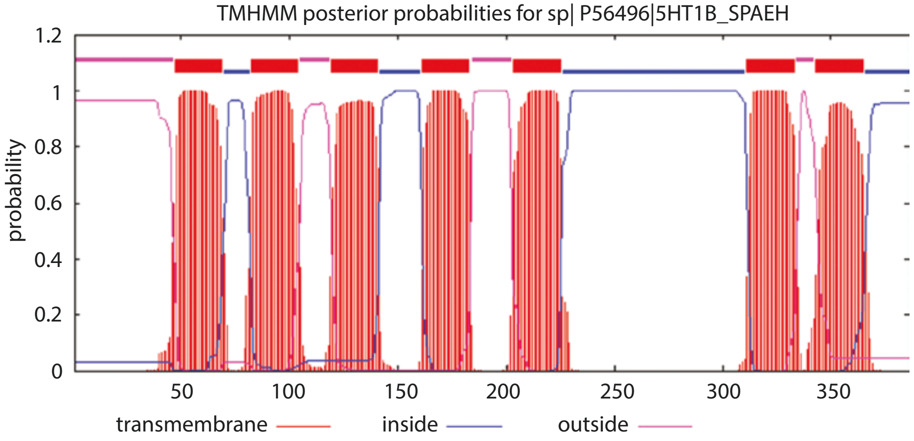
Before the analysis is started it is important to choose the right organism group because the gram-negative bacteria, gram-positive bacteria, or eukaryotes. Figure 5.2 shows the graphical output of the SignalP program for the outer membrane protein C (precursor) from *Salmonella typhimurium* (OMPC-SALTY, P0A263). The C-score stands for cleavage site score, which was trained on the recognition of the cleavage site between signal peptide and the protein sequence, and predicts the cleavage site of SPase I. The maximum C-score occurs at the position of the first amino acid of the mature protein, so one position behind the cleavage site. The S-score, the signal peptide score, is trained on the differentiation of signal peptides and other sequences and has a high value if the corresponding amino acid is part of the signal peptide. Therefore, amino acids of the mature protein have a low S-score. The Y-score (combined cleavage site score) is a geometrical mean of the C-score absolute values and the gradient of the S-score and shows where the C-score is high and the S-score has its inflection point. Analysis of the three scores shows the likely cleavage site between amino acids 21 and 22. In addition, two more values are calculated. The S-mean is the average of the S-scores of all amino acids of the signal peptide. Consequently, if there is a signal peptide, this value should be high. The D-score is the arithmetic mean of the S-mean value and the maximum value of the Y-score. It will also be high if a signal peptide has been predicted.

5.3 Transmembrane Proteins

Biological membranes contain integral proteins that have various functions in the cell, such as acting as cell–surface receptors. Integration into the membrane lipid bilayer is accomplished by hydrophobic interactions between the protein and the nonpolar chain structures of the lipids. The polar head groups of the lipids build hydrogen bonds and ionic bonds with the protein. Integral membrane proteins are therefore always amphiphilic molecules that have both hydrophilic and lipophilic regions. These proteins are orientated asymmetrically in the membrane, i.e., some membrane proteins are only exposed on one side of the membrane, whereas others completely penetrate the membrane and are exposed on both the extracellular and intracellular sides. The latter are called transmembrane proteins. The hydrophobic transmembrane domains are usually formed by α-helices.

The prediction of transmembrane proteins is of great importance for classification and defining function, as described previously for signal peptides. The program TMHMM [tmhmm] of the CBS server in Denmark can predict transmembrane domains. TMHMM is based on a hidden Markov model (HMM) that has been trained to detect hydrophobic transmembrane helices. Furthermore, the program also predicts the orientation of the individual domains in the membrane (intracellular or extracellular) and, accordingly, of the whole protein.

The graphical output of such a prediction with TMHMM is shown in ▣ Fig. 5.3 for the transmembrane domains of the G protein-coupled receptor (GPCR) 5-hydroxytryptamine-1B receptor of the mole rat *Spalax leucodon ehrenbergi* (5H1B-SPAEH). Such GPCRs are integral membrane proteins with typically seven transmembrane domains. In the graph, the probability of a transmembrane helix and its intracellular or extracellular localization is plotted along the amino acid sequence. Additionally, in the upper part of the figure, a schematic representation of the topology is inserted. The graphical representation of the probabilities also allows the recognition of transmembrane helices of relatively low likelihood.



▣ **Fig. 5.3** Graphical output of TMHMM server [tmhmm] at CBS

5.4 Analyses of Protein Structures

As stated earlier, the prediction of a protein 3D structure from an amino acid sequence is currently not feasible and will not be feasible for the foreseeable future. Therefore, experimental methods must be employed to determine protein structures. The two primary approaches are X-ray crystallography and high-resolution nuclear magnetic resonance (NMR) spectroscopy. A third approach using the electron microscope is useful for large proteins. Overall, despite much technological progress, these methods are still very time-consuming and costly, and the successful resolution of a crystal structure is not guaranteed for every protein.

5.4.1 Protein Modeling

A useful and fast method for structure prediction is homology modeling of proteins based on sequence homology. The approach is based on the fact that related proteins within a protein family that have a high degree of amino acid sequence similarity also have similar protein folds (e.g., cysteine proteases of the cathepsin family) (see also ◘ Figs. 5.5 and 5.6). Proteins for which the 3D structure is already known serve as reference proteins or templates. First, the amino acid sequence of the protein to be modeled is compared with the sequence of the reference protein(s) using pairwise or multiple-sequence alignments (in case of several reference proteins). For sequences with identities of more than 70%, the modeled structures can be predicted very accurately. However, for sequences with identities of less than 30%, difficulties with the modeling can arise. The sequence identities of structurally conserved regions (SCRs) are frequently above those of less conserved loops, and both influence the degree of identity of the complete sequence. Interestingly, areas of little conservation are often found at the protein surface and have a comparatively small effect on SCRs, which are found inside the protein where most of the active centers are situated.

To identify SCRs in reference proteins, a structural alignment of the amino acid sequences based on the secondary structure is performed. The sequence to be modeled is then arranged on the oriented templates, and the spatial coordinates of the SCRs are then transferred to the model sequence. The coordinates of the loops are usually taken from similar regions of other protein structures. The spatial orientation of the side chains of individual amino acids in the SCRs is maintained as in the templates. For all nonconserved side chains, the statistically most likely position is taken. The process of homology modeling is completed both by calculations that lead to energy minimization of the model and checking of the structural relevance of the resulting protein model. The SWISS-MODEL Server [swiss-model] of the Swiss Institute of Bioinformatics in Lausanne can be used for the automatic calculation of homology models (Biasini et al. 2014). In the case of proteins with a high sequence similarity, the calculated models are often of high quality.

5.4.2 Determination of Protein Structures by High-Throughput Methods

The number of experimentally determined protein structures stored in the world's only archive for structures of biological macromolecules, the Protein Data Bank (PDB), has grown enormously in recent decades [pdb] (Westbrook et al. 2003). In 1972, PDB contained just one structure, in 1992 the number was approximately 1,000, and by April 2003 it had grown to 20,622. In November 2016, the PDB contained 124,029 structures. This remarkable increase in information can be attributed mainly to the technology process including automatization and high-throughput approaches for solving a structure. The Protein Structure Initiative was one main contributor to this advancement. This initiative was an international scientific consortium of different national initiatives from Japan, North America, and Europe. The aim was nothing less than to structurally solve all of the proteins encoded in the genomes of the most important organisms (archaebacteria, eubacteria, and eukaryotes).

To solve the structures, X-ray structural analyses and NMR spectroscopy were used in a high-throughput format. To decrease the number of protein structures to be experimentally solved, only representatives of the different protein families were examined. The underlying idea was that proteins could be divided into protein families and that sequence similarity usually leads to structural similarity. The conclusion is that the number of different structural folds of proteins found in nature must be limited. One estimate is that between 10,000 and 30,000 protein families exist, and these contain approximately 1000–5000 different protein folds. Of these, approximately 1400 folds are currently known. However, it must be considered that similar protein structures do not inevitably have similar functions and that different protein structures may also perform similar functions. For example, the cysteine proteases are divided into three structurally different groups based on protein folding patterns: the papainlike proteases, the Picorna virus proteases, and the caspases.

To accomplish the ambitious goal of the Protein Structure Initiative, the strategy was as follows:

1. All known protein sequences were grouped into protein families using bioinformatics methods.
2. Representatives of each protein family were produced in sufficient quantity by molecular biological methods.
3. The protein structures of these representatives were experimentally determined using protein crystallography or NMR spectroscopy.
4. All other protein structures of the respective protein families were generated by homology modeling.

Using this procedure a huge amount of new protein fold patterns were identified, thereby making an important contribution to the functional elucidation of all known proteomes. In the meantime, the benefit for modern pharmaceutical drug research was questioned, since most protein structures were solved without function annotation. However, the results will be invaluable in the future. The current initiative called the Structural Genomics Consortium is therefore more focused on the structural solution of diseases-relevant proteins. It should be possible to utilize these structures for structure-based rational drug design and significantly support drug development (Burley and Bonanno 2002).

5.5 Structure-Based Rational Drug Design

From the sequencing of whole genomes and the generation of the corresponding biological information, a modern approach to pharmacological research has been established. To initiate the development of a new drug, a drug target (de facto, a protein) that plays a key role in the disease must first be identified (see also ▶ Chap. 7). After the target's function has been experimentally confirmed (drug target validation), small-molecule chemicals are identified that influence the protein's function in such a way as to alleviate or cure the corresponding disease. The specific inhibition of an enzyme by a chemical inhibitor would be an example.

The overlapping technologies of computer-assisted approaches like bioinformatics, chemoinformatics, and molecular design have become essential components of modern drug discovery efforts. These strategies are indispensable for the identification and validation of drug targets as well as for the screening and design of new small molecules. Also of special importance is the three-dimensional structure of the drug target to allow for structure-based rational drug design. For example, virtual screening, which tests the protein target's interaction with chemical entities in large compound libraries, is an established approach that is incorporated into most discovery strategies. Unlike experiments conducted in the laboratory, virtual screening is automated, being conducted at a computer, and many chemical substances can be tested for their activity spectrum relatively quickly. The most important approaches are pharmacophore-based screening (Wolber et al. 2008) and docking (Kitchen et al. 2004).

The word *docking* is the modern pictorial paraphrase of the lock-and-key concept postulated in 1894 by Emil Fischer (Fischer 1894). The specificity of the receptor–ligand complex is brought about by the geometric and physicochemical complementarity of both. Induced fit is another kind of this hypothesis, where the geometry of the binding site is adapted upon ligand binding. The best known programs in use are DOCK, developed by Irvin Kuntz at the University of California, San Francisco [dock] (Ewing and Kuntz 1996), GOLD from the Cambridge Crystallographic Data Centre [gold] (Jones et al. 1997), FlexX of BioSolveIT GmbH in Sankt Augustin [flexx] (Rarey et al. 1996), and Autodock developed at the Scripps Research Institute [autodock] (Morris et al. 2009).

5.5.1 A Docking Example Using DOCK

With DOCK all possible orientations of a ligand and its receptor can be generated. For example, the protein structure of an enzyme with a clearly defined active center can constitute a typical receptor. The structure of the ligand can originate from a database of chemical molecules such as the Available Chemicals Directory.

In the example shown, the cathepsin L-like cysteine protease of the infectious third-stage larva of the filarial worm *Brugia pahangi* serves as receptor. This enzyme is important for the molting and development of this parasite. The protein structure was generated by homology modeling.

1. The first step is the characterization of the active center (site characterization, ◘ Fig. 5.4). To do this, the molecular surface of the active center is generated first (subprogram MS) and converted to a negative image (subprogram SPHGEN). Overlapping spheres are then fitted into the active center (◘ Fig. 5.5). The center of the spheres will eventually be replaced by the atoms of the ligand.
2. In the second step, a calculation of physical, chemical, and topological parameters is carried out at each nodal point of a space grid (grid calculation) in order to compute a score, which can be either a contact score based on the ligand fit or a force field score.

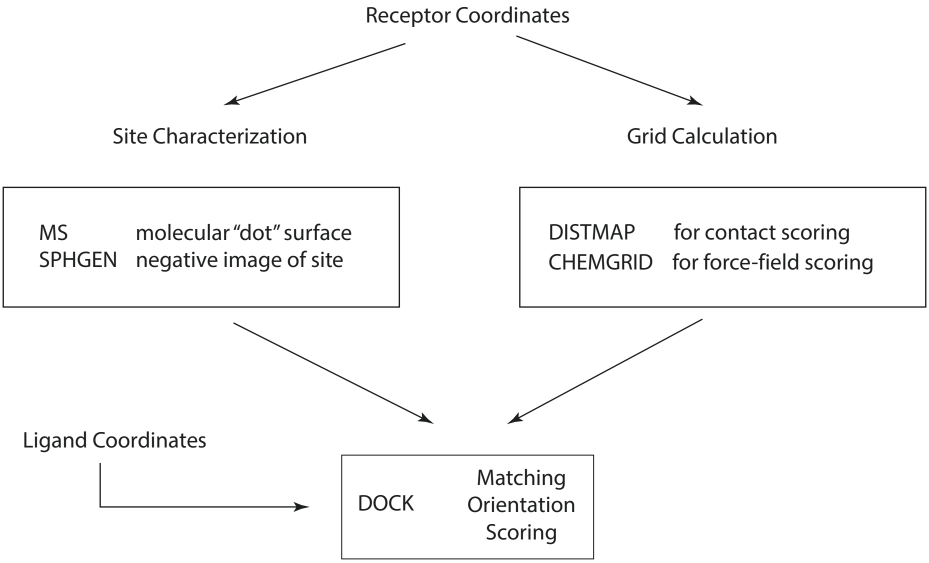

Receptor Coordinates
Site Characterization
Grid Calculation
MS molecular "dot" surface
SPHGEN negative image of site
DISTMAP for contact scoring
CHEMGRID for force-field scoring
Ligand Coordinates
DOCK
Matching
Orientation
Scoring

Fig. 5.4 Schematic representation of mode of operation of program DOCK [dock]

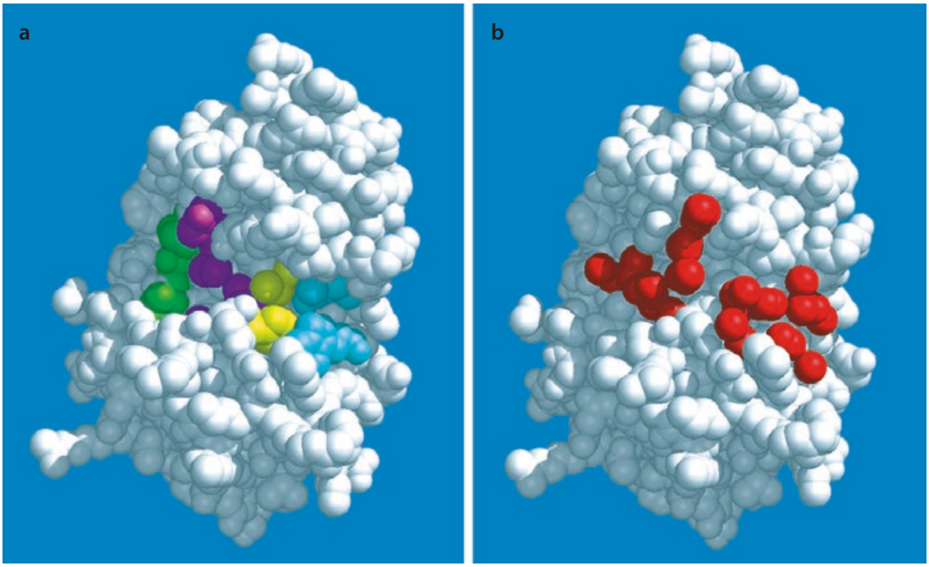
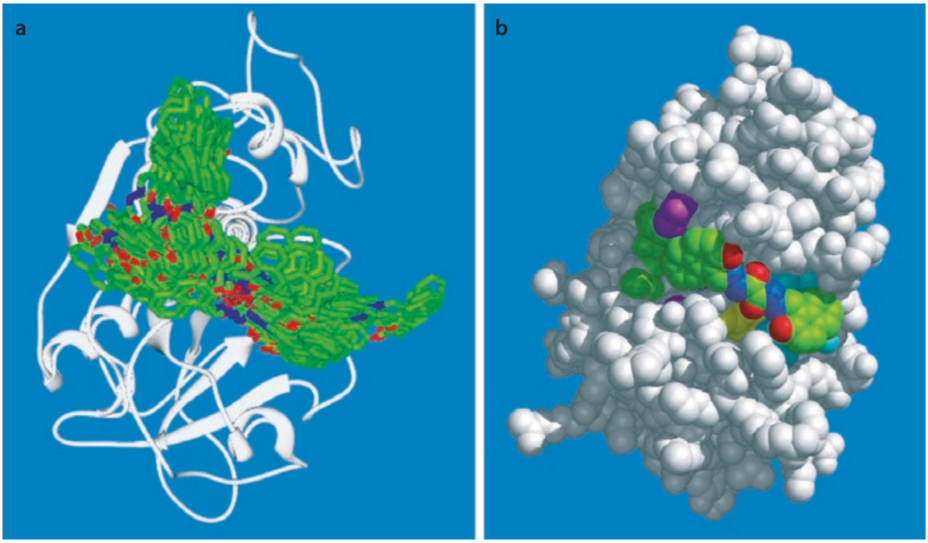


Fig. 5.5 Spherical model of cathepsin L-like cysteine protease of filarial *Brugia pahangi*. The underlying protein structure was generated by homology modeling. **a** The most important amino acids in the active site cleft, which is located between the two main domains of the protein, are indicated in color. The active site cysteine (top) and histidine (bottom) of the catalytic triad are highlighted in yellow. The active site asparagine is hidden in the structure. Important amino acids of the S′ subunit are drawn in cyan, and those of the S subunit are in green and pink. **b** Graphical representation of characterization of catalytic cleft by DOCK program (subprogram SPHGEN). The centers of the overlapping spheres, where later the atoms of the ligands will lie, are represented in red



■ **Fig. 5.6** Model of catalytic cleft of cathepsin L-like cysteine protease of *Brugia pahangi*, into which a chemical compound was modeled using DOCK. **a** The protein is displayed in its secondary structure (ribbon model). In single DOCK mode, all possible orientations of a chemical compound (hydrazide) were generated. All overlapping orientations of this single compound are represented in the figure: carbon, green; oxygen, red; nitrogen, blue. **b** Based on the analysis of the docking experiment described in panel **a**, the most likely orientation of the hydrazide in the catalytic cleft of the cysteine protease is shown. Protein and chemical compounds are represented as spheres. Coloration is similar to those in panel **a** and ■ Fig. 5.5

3. After these calculations, the actual docking can take place. This can be done in two modes, the single DOCK mode or the search DOCK mode. In the single mode, DOCK generates all possible orientations of a single ligand in the active center (■ Fig. 5.6). In the search mode, large databases of chemical molecules are searched. To do this, the best orientation of every ligand is first generated and then saved as a relative score in comparison to all other ligands. The connections with the highest-ranking scores are examined for size, fit, and interaction with the active center. The best compounds can then be experimentally tested for activity in appropriate assays.

For the example of the cysteine protease of *Brugia pahangi* (Lecaille et al. 2002), a chemical database of known cysteine protease inhibitors was searched with DOCK, and hydrazide compounds known to inhibit the cysteine proteases of the parasites Trypanosoma cruzi, Trypanosoma brucei, Leishmania major, and Plasmodium falciparum had very high scores. The binding of the identified hydrazides was then more thoroughly examined in the single DOCK mode to identify the most promising inhibitors (■ Fig. 5.6). Subsequent experiments with the best predicted inhibitors prevented the development of the infectious third-stage larva to the fourth-stage larva (Selzer 2003).

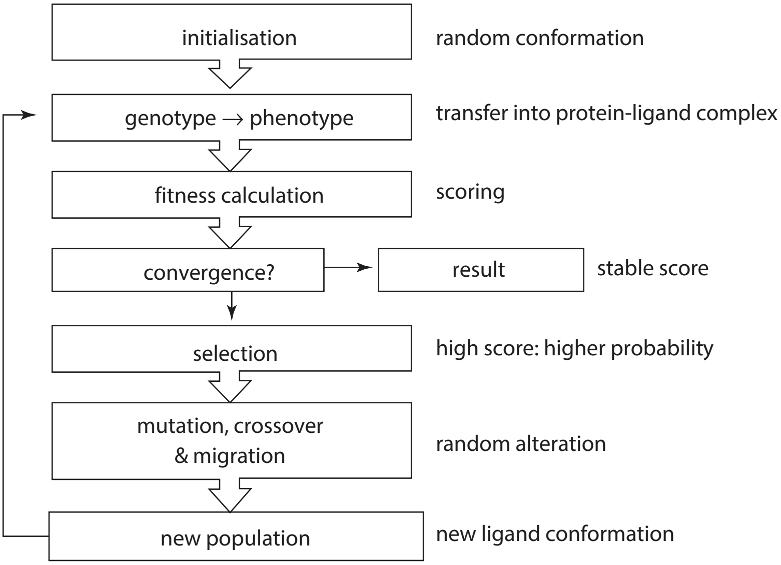


Fig. 5.7 Genetic algorithm used by docking software GOLD

5.5.2 Docking Example Using GOLD

GOLD is another widespread docking software. The ligand conformations in the binding site are calculated using a genetic algorithm that is based on natural genetic evolution (Fig. 5.7). The three-dimensional conformation of a molecule is described via its torsion angles, which are stored as a bitvector (the chromosome). Analogously to nature, this would represent the genotype of ligand conformation. Evolutionary processes are simulated by the mutation of single bits or exchange of single parts of the bitvector of two different chromosomes. Thus, the three-dimensional conformation is changed based on a random alteration of its chromosome. Subsequently, the three-dimensional conformation (the phenotype) will be created based on the chromosome and fitted into the binding site. Each binding pose will be assessed using its fitness based on a scoring function. Only poses with interactions favorable to the protein showing a high fitness will be selected for the next round of the genetic algorithm. These steps are repeated until a stable score is reached.

The protein–protein interaction between thioredoxin reductase (TrxR) and its substrate thioredoxin (Trx) from *Mycobacterium tuberculosis* is a new target for the fight against tuberculosis. The docking software GOLD was successfully used for the identification of the first inhibitors of this protein–protein interaction (Koch et al. 2013). Protein–protein interactions are in general a particular challenge since they are mostly based on the surface of the proteins without any deep binding pocket involved. A detailed analysis of the available X-ray structure revealed an interesting point of targeting. An argine side

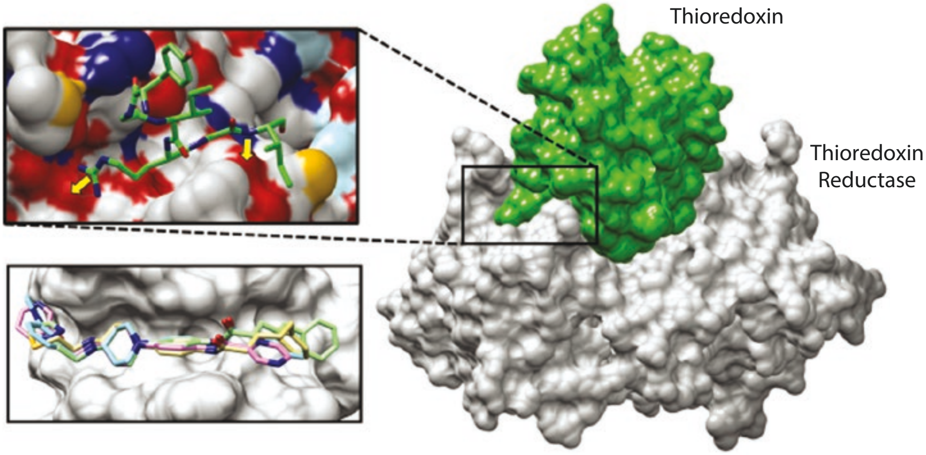


Fig. 5.8 Thioredoxin reductase (gray) and its substrate thioredoxin (green). The top cutout shows the target point for the docking-based virtual screening (yellow arrows: hydrogen bonding interaction). The bottom cutout shows the docking poses of four inhibitors sharing a common scaffold

chain looks out of the globular thioredoxin like an anchor group (Fig. 5.8, top section), together with a hydrophobic cleft and an additional hydrogen binding interaction nearby. GOLD was successfully applied to enrich molecules from a virtual library of 6.5 million compounds that presumably bind to this hydrophobic cleft showing both described hydrogen bonding interactions (Fig. 5.8, bottom section). The 170 best ranked ones were evaluated in a biochemical assay, with 18 molecules showing an inhibition. Overall, this represents a hit rate of 10.5%. Although one could expect a higher hit rate by docking approaches at first sight, this is a remarkable result compared to the possible alternative. Only the X-ray structure was known at the beginning of the study. Thus, all 6.5 million compounds would have been tested to find the first inhibitors. The experimental effort would have been disproportionate in comparison to the 170 tested compounds.

5.5.3 Pharmacophore Modeling and Searches

Knowledge of the 3D structure of a receptor is essential for docking in a virtual screen. In the absence of either a 3D structure or homology model, virtual screening can nonetheless be attempted. As long as some ligands of the receptor are known, virtual screening based on a pharmacophore model can be applied. A pharmacophore model is an abstract concept that considers the interaction potential of a molecule with its target protein and describes the spatial arrangement of the ligand properties that are responsible for binding (Wolber et al. 2008). A superposition of several known inhibitors or active ligands of a protein based on their pharmacophoric properties can be used to construct a pharmacophore model. Possible pharmacophore features are hydrogen bond acceptors and donors, hydrophobic and aromatic systems, and charges (Fig. 5.9). The spatial arrangements of the individual properties are analyzed to retrieve a pharmacophore model or a

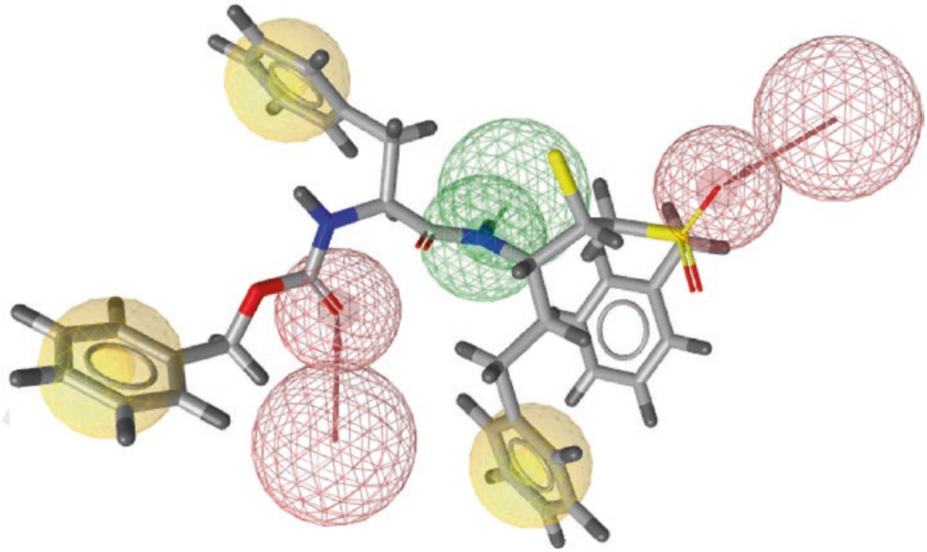


Fig. 5.9 Representation of a pharmacophore model. The pharmacophore features are shown as colored spheres. The shown properties are aromatic (yellow), hydrogen bond acceptor (red), and donor (green). A molecule fulfills this model when the pharmacophore features overlay with the pharmacophore model

pharmacophore hypothesis. By the analysis of virtual libraries based on this pharmacophore model, molecules can be identified that show a similar spatial interaction pattern with a potential similar activity.

All possible 3D conformations of the virtual library must be created for the final screening process since the spatial arrangement of pharmacophore features are fitted in and compared to the pharmacophore model. The overlap is described by a score, and molecules with high agreement can be used as potential ligands for experimental evaluation. The reduced computational time in comparison to docking is the huge benefit of pharmacophore searches. For this reason they are often applied to reduce the size of virtual libraries and filter for subsequent docking. Software for pharmacophore modeling and pharmacophore searches include, for example, MOE Pharmacophore Modeling [moe], Phase [phase] (Dixon et al. 2006), and Ligandscout [ligandscout] (Wolber et al. 2007).

When the 3D structure of a receptor is known, receptor-based pharmacophore modeling can also be used. In addition, protein–ligand complex structures can be used to create a pharmacophore model that contains information about the protein structure and the known ligand (Wolber et al. 2007).

5.5.4 Successes of Structure-Based Rational Drug Design

A frequently asked question is whether such virtual methods actually lead to drugs. The answer is clearly yes. There are more examples than can be listed here where virtual technologies have contributed considerably to the development of drugs. One should keep in mind, however, that the development of a drug is a demanding process that involves many different steps. Rational drug design is only a first step on the long road to a marketable drug.

Dorzolamid (trade name Trusopt, marketed by Merck since 1995), which is used for the treatment of glaucoma, is a carbonic anhydrase inhibitor that originated as the first drug from a program involving structure-based rational design. A second example, Captopril, is a drug that lowers blood pressure whose lead structure was based on a natural substance that inhibits angiotensin-converting enzyme (ACE). Enalapril, another effective ACE inhibitor, is a further development of Captopril. Further examples are the HIV protease inhibitors Saquinavir and Ritonavir (Norvir) from Roche and Abbott, respectively; the tyrosine kinase inhibitor Gleevec from Novartis, which is used in leukemia patients; and the neuraminidase inhibitors Tamiflu, from Roche, and Relenza, from GlaxoSmithKline, which would never have been developed without rational drug design (Klebe 2013).

There are a number of examples where the DOCK program has been used successfully. Particularly impressive have been studies with cysteine proteases. Using DOCK and homology models of the cysteine proteases of *Leishmania major*, small molecules were identified that block these drug target enzymes and stop the development of promastigote and amastigote Leishmania in cell culture without damage to host cells. In a mouse model of leishmaniasis, progression of the infection could be considerably delayed (Selzer et al. 1997, 1999). Similar results in animal models were achieved for cysteine proteases of *Plasmodium falciparum* (Shenai et al. 2002), *Trypanosoma cruzi* (Engel et al. 1998), and *Schistosoma mansoni* (Abdulla et al. 2007). For *Trypanosoma cruzi*, the success of cysteine protease inhibitors has set the stage for clinical trials against Chagas disease (Barr et al. 2005).

Rational design was also used for the development of proteasome inhibitors of parasites. The proteasome is a multicomponent complex of protease, which regulates, for example, important processes of the cell cycle. A detailed analysis of substrate specificity and the protein structure led to selective proteasome inhibitors of *Plasmodium falciparum* (Li et al. 2016). These inhibitors are able to inhibit the parasite's growth in vivo without affecting the host cells. Another proteasome inhibitor is able to inhibit the proteasome of kinetoplastids. All parasites were killed in in vivo studies using a mouse model (Khare et al. 2016).

5.6 Exercises

Exercise 5.1

Explore how many solved protein structures are currently present in the PDB database [pdb].

Exercise 5.2

Find the entry CHER_SALTY/P07801 in the Swiss-Prot database [swiss-prot]. Does this database record contain information about the tertiary structure of the receptor?

Exercise 5.3

View the PDB database record of the receptor from ► Exercise 5.2 (ID 1AF7) and display the molecular structure with one of the PDB visualization programs (NGL Viewer would be a good choice). What information, especially at the structural level (primary, secondary, tertiary structure), is recognizable?

Exercise 5.4

Use NGL Viewer, which should be supported by all current browsers. What display options provide this viewer? Analyze the ligand interaction by selecting the ligand in the *Ligand Viewer* option.

Exercise 5.5

Carry out some secondary structure predictions with the amino acid sequence of the Swiss-Prot database record CHER_SALTY. The necessary programs can be found at ► http://www.expasy.org/proteomics/protein_structure. For example, use the JPred server predicted secondary structure and compare it to the experimentally determined secondary structure.

Exercise 5.6

Does CHER_SALTY have a signal peptide? Give reasons for your assumption. Check the presence of a signal sequence using SignalP [signalp]. Note: *Salmonella typhimurium* is a gram-negative bacterium.

Exercise 5.7

Retrieve the Swiss-Prot database record P41780 from the Swiss-Prot database [swissprot] and repeat ► Exercise 5.6 with this sequence. How does the program SignalP work?

Exercise 5.8

The prediction of transmembrane regions works in a very similar way to the determination of signal peptides. The appropriate program can be found at ► http://www.cbs.dtu.dk/services/. Determine the transmembrane regions of the G-protein-coupled receptor (GPCR) with the Swiss-Prot accession number (AN) Q99527. How many transmembrane regions are detected? Compare this result with a secondary structure prediction for this receptor. Note: In general transmembrane regions are helices.

Exercise 5.9

Perform homology modeling with the Swiss-Prot sequence P29619. To do this, go to the Swiss model page of the Expasy server [swissmodel] and follow the hyperlink *Start Modeling*. Paste the sequence into input field and start building the model with *Build Model*. Save the text file returned by the server with the ending .pdb and open this file with the Swiss PDB viewer. The viewer is available free of charge at the Web site [spdbv]. Tutorials for using spdbv can be found at ► http://www.expasy.org/spdbv/text/main.htm. Another free visualization program is Chimera [chimera].

References

Abdulla MH, Lim KC, Sajid M, McKerrow JH, Caffrey CR (2007) Schistosomiasis mansoni: novel chemotherapy using a cysteine protease inhibitor. PLoS Med 4:e14

Barr SC, Warner KL, Kornreic BG, Piscitelli J, Wolfe A, Benet L, McKerrow JH (2005) A cysteine protease inhibitor protects dogs from cardiac damage during infection by Trypanosoma cruzi. Antimicrob Agents Chemother 49:5160–5161

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42(W1):W252–W258

Biobel G, Sabatini DD (1971) In: Manson LA (ed) Biomembranes. Plenum, New York, pp 193–195

Burley SK, Bonanno J (2002) Structuring the universe of proteins. Ann Rev Genomics Hum Genet 3:243–262

Dixon SL, Smondyrev AM, Rao SN (2006) PHASE: a novel approach to pharmacophore modeling and 3D database searching. Chem Biol Drug Des 67:370–372

Engel JC, Doyle PS, Hsieh I, McKerrow JH (1998) Cysteine protease inhibitors cure an experimental Trypanosoma cruzi infection. J Exp Med 188:725–734

Ewing TJA, Kuntz ID (1996) Critical evaluation of search algorithms for automated molecular docking and database screening. J Comp Chem 18:1175–1189

Fischer E (1894) Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges 27:3189–3232

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748

Khare S, Nagle AS, Biggart A, Lai YH, Liang F, Davis LC, Barnes SW, Mathison CJ, Myburgh E, Gao MY, Gillespie JR, Liu X, Tan JL, Stinson M, Rivera IC, Ballard J, Yeh V, Groessl T, Federe G, Koh HX, Venable JD, Bursulaya B, Shapiro M, Mishra PK, Spraggon G, Brock A, Mottram JC, Buckner FS, Rao SP, Wen BG, Walker JR, Tuntland T, Molteni V, Glynne RJ, Supek F (2016) Proteasome inhibition for treatment of leishmaniasis, Chagas disease and sleeping sickness. Nature 537(7619):229–233

Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3(11):935–949

Klebe G (2013) Drug design. Springer, Heidelberg

Koch O, Jäger T, Heller K, Khandavalli PC, Pretzel J, Becker K, Flohé L, Selzer PM (2013) Identification of M. tuberculosis thioredoxin reductase inhibitors based on high-throughput docking using constraints. J Med Chem 56(12):4849–4859

Lecaille F, Kaleta J, Brömme D (2002) Human and parasitic papain-like cysteine proteases: their role in physiology and pathology and recent developments in inhibitor design. Chem Rev 102:4459–4488

Li H, O'Donoghue AJ, van der Linden WA, Xie SC, Yoo E, Foe IT, Tilley L, Craik CS, da Fonseca PC, Bogyo M (2016) Structure- and function-based design of Plasmodium-selective proteasome inhibitors. Nature 530(7589):233–236

Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J Comput Chem 16: 2785–2791

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786

Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489

Selzer PM (2003) Structure-Based-Rational-Drug-Design: Neue Wege der modernen Wirkstoffentwicklung. In: Lucius R, Hiepe T, Gottstein B (eds) Grundzüge der allgemeinen Parasitologie. Parey, Berlin

Selzer PM, Chen X, Chan VJ, Cheng M et al (1997) Leishmania major: molecular modeling of cysteine proteases and prediction of new nonpeptide inhibitors. Exp Parasitol 87:212–221

Selzer PM, Pingel S, Hsieh I, Ugele B et al (1999) Cysteine protease inhibitors as chemotherapy: lessons from a parasite target. Proc Natl Acad Sci U S A 96:11015–11022

Shenai BR, Semenov AV, Rosenthal PJ (2002) Stage-specific antimalarial activity of cysteine protease inhibitors. Biol Chem 383:843–847

Westbrook J, Feng Z, Chen L, Yang H, Berman HM (2003) The protein data bank and structural genomics. Nucleic Acids Res 31:489–491

Wolber G, Dornhofer AA, Langer T (2007) Efficient overlay of small organic molecules using 3D pharmacophores. J Comput Aided Mol Des 20(12):773–788

Wolber G, Seidel T, Bendix F, Langer T (2008) Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discov Today 13(1–2):23–29

Further Reading

chimera. https://www.cgl.ucsf.edu/chimera/
dock. http://dock.compbio.ucsf.edu/
expasy. http://www.expasy.org
flexx. https://www.biosolveit.de/FlexX/
gold. https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/
ligandscout. http://www.inteligand.com/ligandscout/
moe. https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm
pdb. http://www.rcsb.org/
phase. https://www.schrodinger.com/phase
signalp. http://www.cbs.dtu.dk/services/SignalP/
spdbv. http://www.expasy.org/spdbv/
swiss-model. https://swissmodel.expasy.org/
swiss-prot. http://www.expasy.org/sprot/
tmhmm. http://www.cbs.dtu.dk/services/TMHMM/
uniprot. http://www.uniprot.org