13

In Silico Engineering of Proteins That Recognize Small Molecules

Sushil Kumar Mishra, Gabriel Demo,
Jaroslav Koča and Michaela Wimmerová
CEITEC - Central European Institute of Technology, Masaryk University, Brno,
National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno,
Czech Republic

1. Introduction

The ability of proteins to recognize other molecules in a highly selective and specific manner and to create supramolecular complexes has many biological implications. For example, interactions between receptor-ligand, antigen-antibody, DNA-protein, lectin-sugar are involved in many biologically important processes including transcription of genetic information, enzyme catalysis, transmission of nervous and hormonal signals, host recognition by microbes etc. Therefore, characterizing the structure and energy profile of such supramolecular complexes appears as a key factor in understanding biological function. This may have, in many cases, direct pharmacological consequences. The function of many proteins is driven by reversible binding to small molecules, with either activating or inhibitory effect over the protein's activity. Under these circumstances, it is clear that, in any drug design endeavor, where the goal is to find or build a small molecule that can regulate the function of a protein, it is absolutely essential to understand the stability and behavior of protein-ligand complexes.

However, there is also another kind of an approach to determine protein recognition ability and selectivity mechanisms. It is called protein engineering, and it is based on altering the affinity/selectivity of a protein by substituting some amino-acid residues by other ones in order to identify the most important residues and their specific contribution to the binding activity. Protein engineering is useful not only in the characterization of a protein's binding abilities, but also has applications in bioanalysis and biotechnology. For example, a protein may be engineered (i.e., modified by substituting amino acid residues) to bind specific carbohydrates on the cell surface, and subsequently be used as a marker for diseases characterized by such glycosylation. Another pharmacologically relevant event that can benefit from protein engineering is pathogen/host recognition. In this case, protein engineering may be employed, for instance, to mimic bacterial mutations that lead to multi-drug resistance, to understand their mode of action and to develop new antibacterial drugs. This is certainly a timely issue, as infectious diseases are a leading cause of death worldwide, and they are often connected with a drug resistance. A similar situation occurs in the case of viruses, where the high rate of mutation turns the protein of interest into a continuously moving target, making it tedious to develop drugs or vaccines, e.g., for HIV or influenza viruses.

Protein engineering is typically performed *in vitro*, with *in vivo* consequences and applications. In some cases, it may be very efficient to perform computer modeling and simulations before starting wet laboratory experiments. In such cases we are talking about *in silico* protein engineering, and the goal is to design appropriate mutations in a much faster and cheaper way. In this chapter, we will cover the majority of *in silico* approaches used for protein engineering. The chapter describes not only procedures involved in the *in silico* engineering process itself, but also the description what kind of information is necessary to be able to start *in silico* process. The chapter is composed of several sections. The first describes methods for 3D structure prediction, a necessary step to perform any *in silico* engineering, but not involved in the engineering itself. We further describe various approaches for *in silico* mutagenesis. Afterwards we introduce a number of techniques which enable the prediction of the preferred orientation of the ligand in the binding pocket, as well as the calculation of the binding free energy, again a technique not directly included in protein engineering itself, but necessary to perform it. Some successful examples of *in silico* protein engineering are also given. The whole process is schematically shown in the flowchart in Fig. 1.

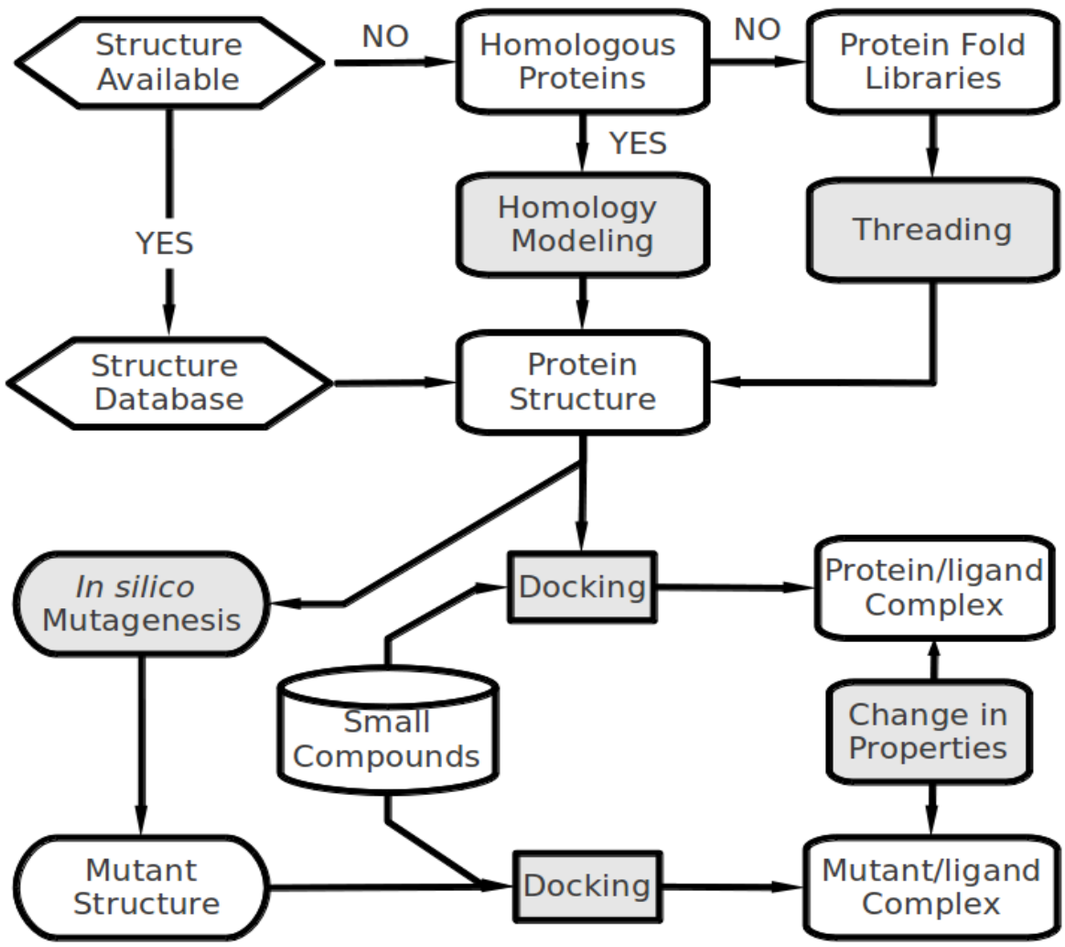


Fig. 1. Flowchart of steps performed within *in silico* protein engineering

2. 3D-Structure as the key prerequisite

A number of proteins that are involved in the cell recognition machinery bind small molecules. In this case, we call these proteins *receptors*, and the small molecules *ligands*. The 3D structure of a receptor is the starting point in the *in silico* protein engineering process. Current experimental methods for protein structure determination are very well established. If the experimental structure is not available, computational approaches are used to model the 3D structure of the receptor.

2.1 Experimental 3D-structure

The 3D protein structure can be obtained by X-ray crystallography or by NMR spectroscopy. Both methods allow to refine the atomic coordinates against experimental structural restraints and constraints. The final 3D model is obtained when the refinement statistics reach relevant global minimum values. The quality of a structure from X-ray crystallography or NMR spectroscopy is defined by experimental data, but the quality of the refined model is based on the interpretation of the model through the personal view of the scientist. In most cases, this freedom in model interpretation is the main source of uncertainty in the results obtained by refining approaches.

2.1.1 X-ray crystallography

The first protein structure determined by X-ray crystallography was solved in the late 1950s. Since that success, over 60 thousand X-ray crystal structures of proteins, nucleic acids and other biological macromolecules have been determined. X-ray crystallography is used to determine the arrangement of atoms in a crystal lattice. The procedure of the 3D structure obtaining is composed of four key steps (see Fig. 2). The crystal under investigation is placed in the way of beams of X-rays, which, upon collision with the crystal, are diffracted in a specific pattern based on the structure of the lattice. The diffraction pattern is used to compute the electron density map of the crystal, from which the mean positions of atoms in the crystal can be determined, together with other information. The resulting electron density map is an average electron density of all the molecules within the crystal. *Structure refinement* refers to the process by which structural models are fit to the information gained from the electron density map. During structure refinement, automated tools for chain tracing, side chain-building, ligand building and water detection are used. The structure refinement continues until the correlation between the diffraction data and the model reach a global minimum (Giacovazzo, 2002).

The atomic positions and their respective B-factors (Debye-Waller factors) can be refined to fit the observed diffraction data. The B-factor, also termed the temperature factor, describes the degree to which the electron density is spread out, accounting for thermal motions and reflecting the fluctuation of atoms about their average positions. Thus, for proteins, the B-factor allows for the identification of areas of large mobility, such as disordered loops, but it can also be the marker of errors in the process of model building (Yuan et al., 2005). The relative agreement of the structure with regard to the experimental data is measured by the R-factor and the "free" R-factor (R-free). The R-free is analogous to the R-factor, which is calculated from a subset (~5%) of reflections that were not included in the structure refinement. The value of R-free is monitored during the whole refinement process, and it prevents any over-refinement and over-interpretation of the data (Brunger, 1992).

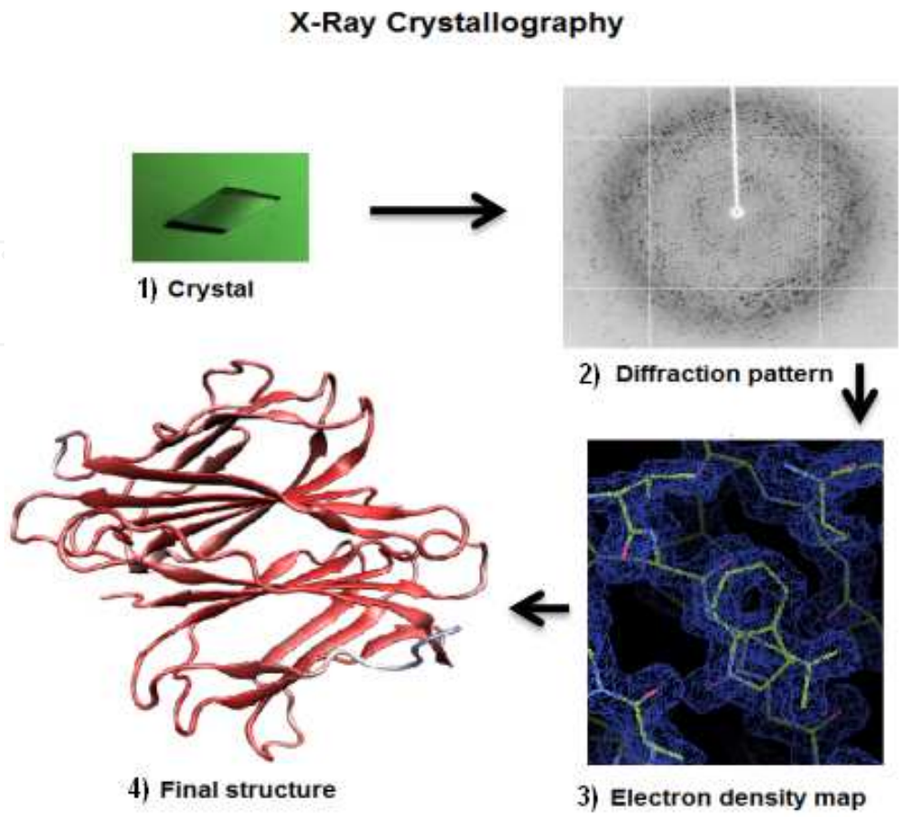


Fig. 2. Four main steps to solve a protein structure by X-ray crystallography: (1) to crystallize the protein, (2) to collect the diffraction, (3) to calculate the electron density map, (4) to refine and validate the model of the structure of the protein

A number of factors contributes to the final quality of an X-ray structure. The first factor relates to the crystal characteristics and its diffraction properties, and is evaluated in terms of resolution. Here, the term *resolution* refers to the level of detail that can be inferred from the electron density map. For proteins, resolutions of less than 2.5 Å are considered meaningful, though the goal is to obtain resolutions of under 1.5 Å, where individual atoms can be clearly pinpointed from the electron density map. Most errors result from highly disordered areas in the electron density maps, like flexible loops of proteins. The electron density of atoms with high residual disorder is smeared in the electron density map, and is no longer detectable. Atoms that give weak scattering (i.e., diffraction of the X-ray beams), such as hydrogen, are normally invisible. Single atoms of protein side chains can be detected multiple times in an electron density map, because of multiple conformations of those respective residues (di Luccio & Koehl, 2011).

2.1.2 NMR spectroscopy

NMR spectroscopy is often the only way to obtain high resolution information on protein dynamics as well as on the protein structure in a solvent. NMR spectroscopy uses the magnetic properties of nuclei that possess a spin. To facilitate NMR experiments, it is

necessary to isotopically label the protein with ^{13}C and ^{15}N (for ^{1}H there is no need to label the protein because this isotope has a natural abundance of 99.9%). The procedure is schematically pictured in Fig.3.

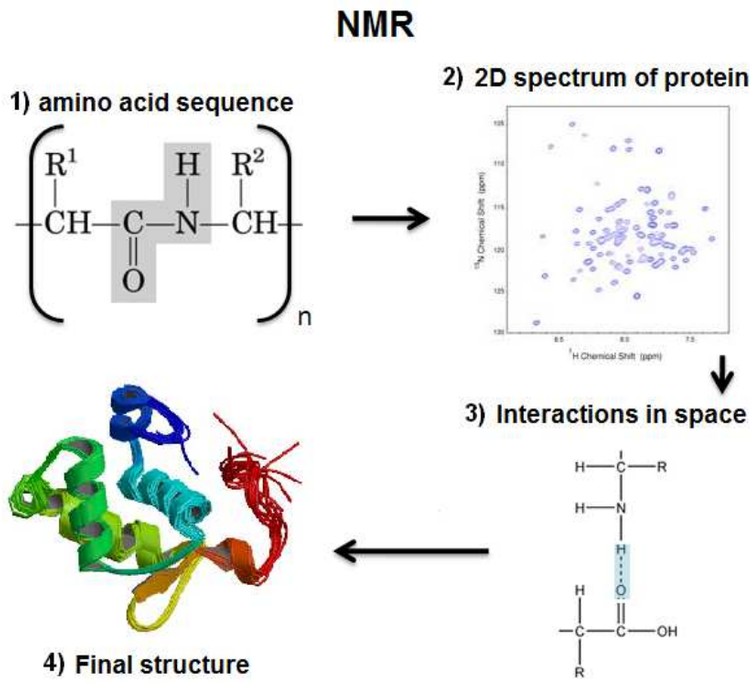


Fig. 3. For solving a protein structure by NMR in solution it is needed: 1) to know the amino acid sequence, 2) to measure the multidimensional spectra 3) to calculate the distances by NOE and J-coupling effects and 4) to refine and validate the 3D structure of the protein

The molecule of interest is placed in a strong magnetic field, and each of these nuclei is characterized by a unique resonance frequency, depending on the electron density of the local chemical environment (chemical shifts), but also on the combination of the local magnetic field and the external field. In the case of proteins, the number of nuclei involved can be large, therefore multidimensional experiments (2D, but also 3D and 4D experiments) are usually performed. The most important method for protein structure determination utilizes NOE (Nuclear Overhauser effect) experiments to measure the distances between pairs of atoms within the molecule that are not connected via chemical bonds (through-space coupling effects). Other NMR experiments are performed in order to measure the distances between pairs of atoms that are connected through chemical bonds (J-coupling). The goal is to assign the observed chemical shifts from multidimensional spectra to their specific atoms (nuclei) in the protein. All the values are then quantified and translated into angle and distance restrains. Most of these restraints correspond to ranges of possible values instead of precise constraints. These restraints are subsequently used to generate the 3D structure of the molecule by solving a distance geometry problem (Wüthrich, 1990, 2003).

The structure determination of macromolecules by NMR spectroscopy shares similarities with X-ray crystallography in terms of possible sources of errors. The errors in an NMR structure can result from an improper experiment setup, as well as from the human

misinterpretation of the experimental data (Saccenti & Rosato, 2008). Molecular modeling techniques are used to generate a set of models for the protein structure that satisfy the obtained experimental restraints, as well as standard stereochemistry. Analogously to X-ray methods, the quality of NMR measurements affects the quality of the structures. The value of the root mean square (RMS) difference between each model and a "mean" structure defines the precision of a set of models for a protein structure. The quality of each model is evaluated by the number of the experimental restraints violations in the final model.

2.2 Homology modeling

Despite significant progress in X-ray crystallography and NMR spectroscopy, the structures of many biotechnologically and therapeutically relevant proteins remain undiscovered for various reasons. In such a case, homology modeling can be used to obtain their 3D structure. Homology modeling is a purely computational procedure that consists of building a protein model using a structural template, normally coming from proteins with a known structure. The procedure is composed of four key steps as seen in Fig. 4.

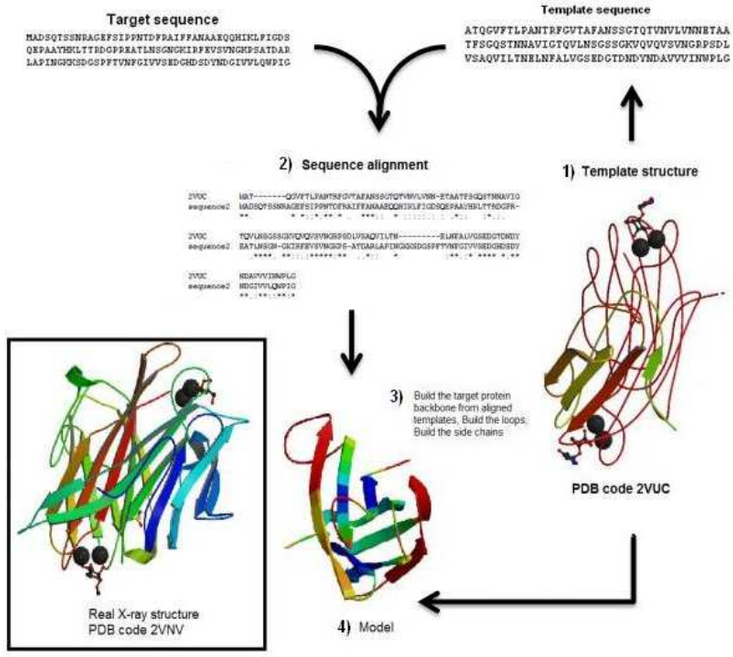


Fig. 4. Homology modeling consists of: 1) Identification of the template, 2) Alignment of the target sequence with the template sequence, 3) Building the target protein backbone, loops and side chains and 4) Refining and evaluating the final model.

Template selection and sequence alignment

An initial step for comparative modeling is to check whether there is any protein in the current PDB database having a similar sequence as the protein of interest. If so, the structure of this protein will be used as a template. The search for the template has to proceed using a sequence comparison algorithm that is able to identify the global sequence similarity (i.e., the degree to which the sequence of amino acids is conserved in the protein under

investigation compared to the template protein). Homology modeling of a target protein sharing over 30% sequence identity with its template is expected to generate structural models whose accuracy is close to that of an experimental structure, however, Roessler and coworkers showed that even proteins sharing 40% of sequence identity can display different folds (Roessler, 2008).

The sequence of the protein with unknown structure is aligned against the sequence of the template protein, meaning that the sequences are arranged in such a way that the regions which contain the same amino acids in both proteins are superimposed. Then the C_α coordinates of the aligned residues from the template are copied over to the target protein in order to form the skeletal backbone (Nayeem et al, 2006). Commonly used alignment techniques are: standard pairwise sequence alignment, where only 2 sequences are compared at a time, or multiple sequence alignment, where more sequences are compared at a time and which is generally used when the target and template sequences belong to the same family. There are complex sequence alignment algorithms that optimize a score based on a substitution matrix and gap penalties. Most errors are caused by the sequence alignment technique. Errors appear frequently in the loop regions between secondary structures, as well as in regions where the sequence similarity is low. Structural alignment techniques are also available, which attempt to find areas of structural similarity between proteins. Recent techniques aim to use as much information as possible while performing the sequence alignment (amino-acid variation profiles, secondary structure knowledge, structural alignment data of known homologs) (Nayeem et al., 2006; Zhang, 2002).

Loop building

Loops participate in many biological events and contribute to functional aspects such as enzyme active sites formation or ligand-receptor recognition. The flexible nature of loops causes problems in the prediction of their conformation. Databases of loop conformations or modeling by *ab initio* methods are used in order to determine the proper structure of loops. In the database approach, a library of protein fragments is scanned for fragments whose length matches to the corresponding lenght of the modelled loop (for short loops) (di Luccio & Koehl, 2011; Zhang, 2002). The *ab initio* loop prediction approach relies on a conformational search guided by various scoring functions and is used for longer loops (Olson et al., 2008; van Vlijmen et al., 1997).

The side-chain positioning problem

Most of the side-chain positioning methods are based on rotamer libraries with discrete side-chain conformations. Rotamer libraries contain a list of all the preferred conformations of the side-chains of all twenty amino acids, along with their corresponding dihedral angles (Lovell, 2000). Side chain prediction techniques choose the best rotamer for each residue of the protein based on a score that includes both geometric and energetic constraints (combinatorial problem). The combinatorial problem is solved by heuristic techniques such as mean field theory, derivatives of the dead-end elimination theorem or Monte Carlo techniques (Vasquez, 1996).

Refinement and validation of the final model

When determining the structure of a protein by homology modeling, the last step is refining the model. However, it was shown that refining a structural model by energy minimization

only (i.e., without experimental constraints) many times leads to structures that are different compared to those obtained by X-ray crystallography. To avoid such problems, several approaches can be applied including evolutionary derived distance constraints (Misura et al., 2006), the combination of molecular dynamics and statistical potentials (Zhu et al., 2008), adding a differentiable smooth statistical potential (Summa & Levitt, 2007) or considering the solvent effects (Chopra et al., 2008).

For the model validation step, scoring functions are used. These are functions based on statistical potentials, local side-chain and backbone interactions, residue environments, packing estimates, solvation energy, hydrogen bonding, and geometric properties. The validation of models can also come from experiments, and further later experimental constraints/restraints can be used to improve the accuracy of the respective models (di Luccio & Koehl 2011).

Generally, the quality of the homology model is dependent on the quality of the sequence alignment and of the template structure. The presence of alignment gaps (commonly called indels) in the target but not in the template complicates the model building process. In addition, it's very hard to deal with the gaps in the template structure (e.g, caused by the poor resolution of an X-ray structure). At 70% sequence identity between the model and the template, the root mean square deviation (RMSD) between the coordinates of the corresponding C^{α} atoms is typically ~1–2 Å. The RMSD can rise to 2–4 Å at 25% sequence identity. The errors are significantly higher in the loop regions, because of the increased flexibility in these areas, both in the target, as well as in the template. Errors in side chain packing and positioning increase with decreasing amino acid sequence identity, and are caused also by the fact that most side chains can exist in several conformations. These errors may be significant, and they imply that homology models must be utilized carefully. Nevertheless, homology models can be useful in reaching *qualitative* conclusions about the biochemistry of the query sequence (conserved residues can stabilize the folding, participate in binding small molecules or play a role in the interaction with another protein or nucleic acid) (di Luccio & Koehl 2011). The state of the art in homology modeling is assessed in a biannual large-scale experiment known as the Critical Assessment of Techniques for Protein Structure Prediction, or CASP. A particularly interesting example is provided by the application of homology modeling to virtual screening for GPCR (G-protein coupled receptor) antagonists (Evers & Klabunde, 2005).

Online portals, such as the Protein Structure Initiative (PSI) model portal (http://www.sbkb.org), or the Swiss-Model Repository (http://swissmodel.expasy.org), bring to the community a large database of models. The PSI model portal ((http://www.proteinmodelportal.org) currently provides 22.3 million comparative protein models for 3.8 million distinct UniProt entries with relevant validation data.

A variety of software is currently in use for homology modeling of protein structures:

GeneMine: Homology modeling in GeneMine (Lee & Irizarry, 2001) uses SegMod, a segment match modeling protocol (Levitt, 1992). The target sequence is divided into short segments. Corresponding structural fragments are taken from a structural database and then matched to the sequence. The fragments are then fitted onto the framework of the template structure. The program generates 10 independent models, from which an average model is constructed and stereochemically refined to minimize conformational repulsion.

DS MODELER: The protein homology modeling program DS MODELER (Accelrys Software Inc.) includes the software tool MODELLER (Sali & Blundell, 1993). MODELLER makes structure predictions based on distance restraints obtained from the template, from the database of crystal structures in the PDB, and from a molecular force field. Loops are generated *de novo*, by a process that incorporates knowledge-based potentials from known crystal structures.

ICM: The homology modeling option in ICM (Abagyan & Batalov, 1997) is completely automated. The template is used for matching the backbone, as well as the side chain conformations for the residues that are identical to the template. Loops are inserted from conformational databases with matching loop ends. The non-identical side chains are given the most preferred rotamer, and then optimized by torsional scan and minimization.

SWISS-MODEL: SWISS-MODEL is an automated protein structure homology modeling server accessible from the ExPASy Web server (Schwede et al., 2003). The input for SWISS-MODEL is a sequence alignment and a PDB file for the template. The homology model is constructed using the ProModII program (Peitsch, 1995). Model construction includes backbone and side chain building, loop building, validation of the quality and of the packing of the model. The model coordinates are returned in PDB format.

2.3 Threading

Threading is used to model the structure of a protein when no homologs with a known 3D structure are available. Protein threading is based on the idea that there is a limited number of different folds in nature (approximately 1000), and thus a new structure has a similar structural fold to those already deposited in the PDB. The threading approach is a specialized sub-class of fold recognition. It works by comparing a target sequence against a library of potential fold templates using energy potentials and/or other similarity scoring methods. The template with the lowest energy score (or highest similarity score) is then assumed to best fit the fold of the target protein. The procedure is composed of three key steps shown in Fig. 5.

Threading improves the sequence alignment sensitivity by introducing structural information (the secondary or tertiary structure of the targets) into the alignment. For example, some amino acids are preferred in helical secondary structure, some can appear more frequently in hydrophobic environments, *etc*. This different behavior of amino acids produces different secondary and tertiary structures of proteins, depending on what environment they are exposed to.

The earliest threading approach was the '3D profiles' method (Luthy et al., 1992), in which the structural environment at the position of each residue of the template is classified into 18 classes, based on the position status, local secondary structure and polarity.

Frequently used threading methods are based on the Profile Hidden Markov Model method (HMM) (Durbin, 1998). All the sequences in the database are clustered into a set of families. In an HMM algorithm, the target is represented by the predicted secondary structure, while the template structures are represented with the template's secondary structure patterns. The majority of current threading methods are based on *residue pairwise interaction energy* methods, where, in each step of the threading procedure, the alignment score is calculated by adding up all the pairwise interaction energies between each target residue and the template residues surrounding it.

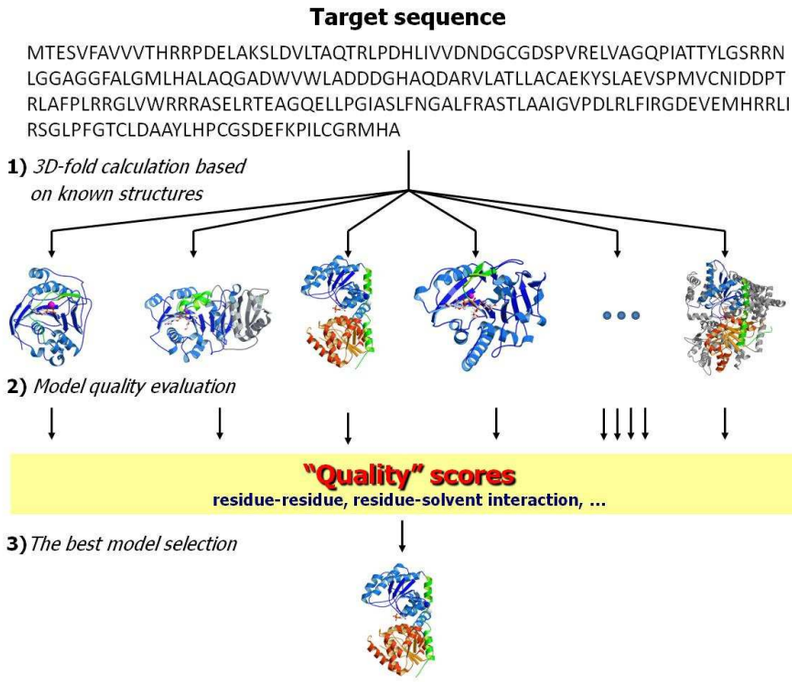


Fig. 5. Three main steps of threading: 1) The construction of a structure target database based on templates, 2) Calculation of the quality of each model and 3) Selecting the best model

Threading methods are not able to give a good sequence–structure alignment. The first reason is that the structure information has many approximations. Most of the threading methods use a '*frozen*' approximation. It means that the target residues are in the same environments as the template residues if they belong to the same structural fold. But, especially in loop regions, two homologous structures can have slightly different environments. Therefore, only conserved regions are used in threading (Madej et al., 1995).

A variety of threading software is available:

GenTHREADER is a fast and powerful protein fold recognition method (Jones, 1999a). It is used to make structural alignment profiles in the construction of the fold library. PSI-BLAST (Position-Specific Iterated - Basic Local Alignment Search Tool, Altschul et al., 1997) profiles, bidirectional scoring and secondary structures predicted by PSIPRED (Jones, 1999b), have also been incorporated into the modified protocol. Because of these implementations, the sensitivity and the accuracy of alignments is increased (McGuffin & Jones, 2003). New implementations for structure prediction on a genomic scale and for discriminating superfamilies from one another were added recently (Lobley, 2009).

3D-PSSM (Kelley et al., 2000) is using PSIPRED to predict the secondary structure of target proteins, and PSI-BLAST for sequence-profile alignments. The target profiles are aligned against 3D position-specific scoring matrices (PSSMs), which are generated for the templates within the fold library. For each template, PSI-BLAST is used to generate an initial 1D sequence based PSSM, which is then further enhanced using solvation potentials, secondary structures and structural alignments, resulting in a 3D-PSSM.

Phyre2 (Kelley & Sternberg, 2009) is a major update to the original Phyre server. It is designed to predict the 3D structure of a protein from its sequence. Phyre2 uses the alignment of hidden Markov models via HHsearch (Söding, 2005) in order to significantly improve the accuracy of the alignment, as well as the rate of detection of homologous regions. For regions that are not detectable by homology, *ab initio* folding simulations called Poing are used (Jefferys et al, 2010).

3. *In Silico* mutagenesis of proteins

The ultimate goal of protein engineering is to design a protein with novel properties, starting from existing proteins. Protein engineering in the field of recognition has been particularly successful in changing ligand specificity and binding affinity. Consequently, we are interested in changing the structure of a macromolecule in a predetermined way, such that we can affect its recognition ability. During the last years, the availability of computational and graphical tools, which allow to display and explore the three dimensional structures of proteins, has made *in silico* mutagenesis easier and more feasible.

Basically, two approaches are available - mutation of a single, or of multiple residues.

3.1 Performing *in silico* mutagenesis

The 3D structure of a protein molecule is generally stored as a text file which contains information about the chains, residues, atoms and atom types, atomic coordinates and their occupancy. Performing *in silico* protein mutagenesis basically means changing the lines of the text that encode the information about the residue being mutated, followed by a set of additional operations meant to properly integrate the mutated residue into the structure.

The mutation of one residue to another does not change anything in the backbone atoms. In addition, the protein side chains all start by the β carbon atom, which is the same for all the amino acids except for the glycine. Therefore, the single amino acid mutation is straightforward, since only the side chain atoms need to be changed. The most critical step is to check for steric clashes that may occur, especially when an amino acid with a short side chain is mutated into another one having a longer side chain. Moreover, the new amino acid may adopt several side chain orientations. This problem is handled using the concept of rotamers, which are defined as low energy side-chain conformations, and are sampled according to their occurrence in proteins. Computational chemistry tools are able to include all the possible side chain conformations by using rotamer libraries. Several molecular modeling platforms facilitate single point mutation using the concept of rotamers.

Some of commonly used software packages to perform single point or multiple point mutations at selected positions:

Swiss-Pdb Viewer: an application that allows to analyse several proteins at the same time (Guex & Peitsch, 1997). The proteins can be superimposed in order to deduce structural alignments and compare their active sites. Swiss-Pdb Viewer allows to browse a rotamer library for amino acids side chains. Amino acid mutations, H-bonds, angles and distances between atoms are easy to obtain.

Pymol: an open-source molecular visualization system (Schrodinger LLC). It can produce high quality 3D images of small molecules and biological macromolecules, such as proteins.

PyMol has a mutagenesis wizard to perform mutations. Several side chain orientations (rotamers) are possible. The rotamers are ordered according to their frequency of occurrence in proteins.

MODELLER: contains the routine 'mutate_model', which allows *in silico* side chain replacement, as well as modeling the final structure of the mutated protein. The routine introduces a single point mutation at a user-specified residue, and optimizes the mutant side chain conformation by conjugated gradient and a molecular dynamics simulation (Sali & Blundell, 1993).

Triton: a graphical interface for computer aided protein engineering. It implements the methodology of *in silico* site-directed mutagenesis to design new protein mutants with required properties, using the external program MODELLER mentioned above. The program allows to perform the one-, two- or multiple-point amino acid substitutions in a very user-friendly and automated way (Prokop et al, 2008). Output data can be easily visualized, written or organized as input files for any of the other computational chemistry modules that Triton interfaces. Routines to study enzyme kinetics and protein/ligand binding are available.

3.2 Alanine scanning mutagenesis

Alanine scanning mutagenesis is a method usually used to determine the contribution of a particular residue to protein function by mutating that residue into alanine. Alanine scanning involves substituting of a larger group of atoms with a smaller one. Alanine is the residue of choice because it removes the side chain beyond the β carbon of the amino acid in question, and, most importantly, because it does not alter the main-chain conformation (Wells, 1991). Additionally, it does not impose extreme electrostatic or steric strain in the system. Glycine would also cancel the contribution of the side chain, but could introduce conformational flexibility into the protein backbone, and therefore is not commonly used.

Alanine-shaving is the process of making multiple simultaneous alanine mutations and can be helpful, e.g., in investigating the cooperativity between side chains (Bogan & Thorn, 1998). Cooperativity can be detected by multiple mutation cycles (Carter, 1986), in which the free energy change caused by the simultaneous mutations at selected residue positions in a protein is compared with the sum of the free energy changes associated with single mutations at each of the selected positions. This technique has also been used experimentally (Bogan & Thorn 1998).

4. Qualitative and semi-quantitative approaches to evaluate the recognition ability of proteins

Molecular recognition can be viewed as the ability of a certain biomacromolecule to interact preferentially with a particular target molecule. A necessary prerequisite for any *in silico* protein engineering approach is the ability to evalute how strong the recognition is. In biological systems, the process of recognition, governed by non-covalent interactions, results in the formation of a complex, where one biomacromolecule interacts with another biomacromolecule or a small molecule. Modern computer modeling and simulation methods, such as docking or free energy calculations, make it possible to study the molecular recognition process between two molecules *in silico*. Evaluation of the recognition

ability of biomacromolecules is performed in two steps: (i) docking the small molecule into the biomacromolecule and (ii) analyzing the interactions and factors that determine the binding affinity.

4.1 Principles of molecular docking

Molecular docking is a widely-used computational tool for the study of molecular recognition, which aims to predict the preferred binding orientation of one molecule to another when bound together in a stable complex. Docking can be performed between two proteins, a protein and a small molecule, a protein and an oligonucleotide or between an oligonucleotide and a small molecule. We use the terms *receptor* and *ligand* to describe the role of binding partners in docking. *Receptor* denotes the system we are docking to (most commonly a protein), while *ligand* denotes the molecule being docked (drug-like compounds, peptide, carbohydrate, etc.). The docking product is commonly referred to as the *complex*. Inside the complex, the position of the ligand relative to the receptor is called the *binding mode*. The space within the receptor where binding modes are explored is commonly known as the *search (or grid) space*.

As already mentioned, receptor-ligand docking programs usually run in two primary parts. The first stage is *searching the grid space* and it leads to the generation of possible binding modes of the ligand within the predefined search space in the receptor. The second stage of docking is *scoring*, and it refers to the process of quantifying the binding strength of each mode of binding using a function called a *scoring function*. We describe each stage in the following pages.

4.2 Receptor site characterization

In the process of docking, the first issue is where to dock the ligand, i.e., how to define a search space on the receptor where the search will be performed. If the 3D structure and the binding site of the receptor is known, the search space is defined within and around this binding site. However, it can happen that the 3D structure of the receptor has not been solved, or there is no experimental evidence indicating a possible region for the ligand binding. In this case, it is recommended to do a prior identification of the binding site by using specialized tools such as PASS (Brady & Stouten 2000), Q-sitefinder (Laurie, 2005), ICM Pocketfinder (An J, 2005) etc. The ligand binding site prediction itself is a complex and tedious problem, thus we will not discuss the details here (for further reading see: Huang & Zou, 2010; Yuriev et al. 2011).

If no prior identification of the binding site is done, it is indeed possible to define the search space around the whole receptor. This approach is known as *blind docking*. A library of ligands is docked into the receptor in order to get an idea of the potential binding regions. The reliability of the blind docking results highly depends on the correct prediction of the binding regions, and represents a compromise between speed and accuracy.

4.3 Sampling protein and ligand conformational flexibility in docking

The main docking operations focus on the ligand. However, during the docking, several preliminary assumptions need to be made about the receptor flexibility. About 85% of proteins undergo conformational changes upon ligand binding, mainly movements in the essential binding site residues (Najmanovich et al., 2000). Therefore, performing accurate

molecular docking is quite difficult, because of the many possible conformational states of both the biomacromolecule, and the ligand flexible areas. Depending on how conformational flexibility is handled during the docking, we distinguish between two classes. The *rigid body docking* method handles both binding partners as rigid bodies. The bond angles, bond lengths and torsion angles of the docking partners are not modified at any stage of the docking. By contrast, in *flexible docking* procedures, binding partners are considered as flexible molecules. This kind of procedure allows the specified atom or group of atoms to acquire the preferred position upon binding. Flexible docking is further categorized into two types: *flexible ligand* docking, where only the conformation of the ligand changes during the docking, and *flexible receptor docking*, where both the conformation of the ligand and the conformation of the receptor can change.

4.4 Sampling conformational and configurational space

Search space where we sample the structural arrangement of two molecules without changing the conformation of any of the molecule is called configurational search space. This term can be used for search space on rigid docking. Whereas in flexible docking, we search for the configurations of the system with two molecules, each of them being able to adopt several conformations. The configurational and conformational search is done via a set of algorithms that sample all the desired degrees of freedom of the ligand in order to find the correct binding mode. The set of operations performed to improve a binding mode is often referred to as *optimization*. Optimization is a difficult problem in docking, because it requires successful conformational search combined with an effective global sampling across the entire range of possible docking orientations.

There are basically three general categories of such algorithms, based on shape matching, systematic search and stochastic search, respectively.

Shape Matching is an approach based on the geometrical overlap between two molecules. The algorithm first generates a "negative image" of the binding site starting from the molecular surface of the receptor, which consists of a number of overlapping spheres of varying radii. The ligand is placed into the binding site using the surface complementarity approach, i.e., the molecular surface of the ligand has to attain maximum close surface contacts to the molecular surface of the binding site of the protein. To do this, ligand atoms are matched to the sphere centres of the negative image. The ligand can then be oriented in the binding site by performing least squares fitting of the ligand atom positions to the sphere centres. The degree of shape complementarity is measured by a certain score function. Maximizing this score function leads to the docked configuration. Note that this is not the function used in the second docking stage, though that one is also referred to as score or scoring function. Examples of docking programs which are based on this approach are DOCK (Kuntz et al., 1982), FRED (McGann et al., 2003) and MS-DOCK (Sauton et al., 2008).

Systematic Search algorithms try to explore all the conformational degrees of freedom of the ligand and combine them with the search on the system with the receptor. Depending on the way how the search is carried out, there are three main subclasses of systematic search algorithms.

A-*Systematic or pseudosystematic search*, where a huge number of poses are generated by rotating all the rotatable bonds by a given interval (in degrees). These poses are then filtered

by using some geometrical and chemical constraints. The remaining poses are subjected to more accurate optimization. This hierarchal sampling method is currently used by the Glide (Friesner et al., 2004) and FRED (McGann et al., 2003) docking programs.

B-*Fragmentation methods* divide the ligand into small fragments (both rigid and flexible). First, a rigid core fragment is placed into the active site. Then, the more flexible fragments are sequentially linked by covalent bonds by using the "place-and-join" approach. Currently, docking programs like LUDI (Böhm, 1992), DOCK (Ewing & Kuntz, 1997), FlexX (Rarey et al., 1996) and eHiTs (Zsoldos et al., 2006) provide this methodology.

C-*Database or conformational ensemble* methods use an ensemble of pre-generated ligand conformations to deal with ligand flexibility, which is then combined with a search for proper receptor/ligand orientation. Databases or libraries of conformations can be generated within the docking program or separately, using other programs such as OMEGA (OpenEye Scientific, NM). FLOG (Miller et al., 1994) is a typical software using this methodology, but some other programs like MS-DOCK (Sauton et al., 2008) and Q-Dock (Brylinski & Skolnick, 2008) also offer this approach.

Random or ***stochastic*** methods are also available. They attempt to sample the space by making random changes to the receptor/ligand system. Whether a geometry change is accepted or rejected is decided using a predefined probability function. This may result in non-reproducible results, even if the docking is repeated with the same parameters. There are mainly four types of stochastic search algorithms.

A. Monte Carlo (MC) is used for a large set of optimization problems, ranging from economics, mathematics to nuclear physics or even regulating the flow of traffic. In docking, the ligand is first placed into the binding site of the receptor, and this binding mode is scored. A new geometry is generated by applying random changes to the rotatable bonds or the position of the ligand with respect to the receptor. The new binding mode is then scored. If the score of the new binding mode is better than that of the old one, this change is accepted. Otherwise, a probability (P) to accept the change is calculated as $P \approx \exp(-\Delta E/K_bT)$. Here ΔE is the change in score, K_b is Boltzmann's constant and T is the absolute temperature of the system. A random number (r), between 0 and 1, is generated, and if $r < P$, the change is accepted. After such an evaluation, another random change is applied to the ligand and the whole procedure is repeated until a reasonable number of orientations is obtained. AutoDock (Morris et al. 1998), ICM (Abagyan et al. 1994) and QXP (McMartin & Bohacek, 1997) are key examples of programs that use MC-based optimization procedures.

B. Genetic Algorithms (GA) are based on ideas derived from natural evolution, such as mutation, crossover, inheritance and selection. To solve the optimization problem, GAs simulate the survival of the fittest among individuals over consecutive generations. Each geometry of the ligand with respect to the protein is defined by a set of state variables called genes. Genes describe the translation, rotation and orientation of the ligand. A full set of a ligand's state variables is referred to as the genotype, whereas the phenotype is represented by the atomic coordinates. Genetic operations such as mutation, crossover, inheritance and selection are applied to the population until the fitness criterion is fulfilled.

Some of the most popular programs like AutoDock (Morris et al., 1998), GOLD (Jones et al., 1995, 1997), and Lead finder (Stroganov et al., 2008) include GA or hybrid approaches to find the optimal orientation of the ligand.

C. Tabu search (TS) is a meta-heuristic approach where a local search is combined with storing a list of previously considered geometries, along with a probability criterion, which ensures that only a new geometry will be sampled further. A random change is only accepted if the RMSD between the new conformation and any of the previously sampled geometries is greater than a threshold. The programs PRO_LEADS (Baxter et al., 1998) and PSI-DOCK (Pei et al., 2006) are TS based software.

D. Particle Swarm optimization (PSO) is one of the evolutionary computational techniques inspired by the social behaviour. SO exploits the population of individual to probe the premising region of search space. The population is called *swarm* and the individuals are called *particles*. These algorithms maintain a population of geometries by modeling swarm intelligence, a concept referring to the collective behaviour of otherwise fully independent particles. A number *particles* is randomly set into motion through this space. At each iteration, they observe the fitness of themselves and their neighbours and emulate successful neighbours (those whose current position represents a better solution to the problem than theirs) by moving towards them. The major advantage of PSO, compared with GA, is its relative simplicity and quick convergence. Examples of docking programs that use swarm optimization are SODOCK (Chen et al. 2007), Tribe-PSO (Chen et al., 2006), PSO@AutoDOck (Namasivayam & Günther, 2007).

4.5 Scoring ligand poses

Once a reasonable set of receptor/ligand geometries has been generated, ranking these modes is the second critical aspect of the docking procedure. To recognize the true binding modes from all the geometries, the binding affinity is scored using scoring functions, i.e., each binding mode is analysed by a set of equations and compared to the other binding modes. If the search algorithms predict a "correct" binding mode but the scoring function fails to rate this as a top scoring orientation, then the suggested output will be a false negative binding mode. Therefore, scoring functions should be able to distinguish between a true binding mode and all other modes explored. However, using a rigorous scoring function for several hundreds of binding modes is computationally expensive. Hence, computationally feasible empirical scoring functions are commonly used by all available docking software. Numerous scoring functions developed and evaluated so far can be grouped into three basic categories.

A. Force field based: A force field is a way to express the potential energy of the system by using a mathematical function and a set of parameters. A basic functional form of a force field encapsulates both bonded terms (between atoms that are linked by a covalent bond) and non-bonded terms (also called " non-covalent "). Non-bonded terms describe van der Waals and long range electrostatic forces. The generic equations (1-3) for force fields such as in AMBER (Weiner & Kollman, 1981) or CHARMM (Brooks et al., 1983), are expressed as:

$$V\left(r^N\right) = V\left(r^N\right)_{Bonded} + V\left(r^N\right)_{Non-bonded} \tag{1}$$

$$V\left(r^N\right)_{Bonded} = \sum_{bonds} \frac{k_i}{2}\left(r_i - r_{eq}\right)^2 + \sum_{angles} \frac{k_\theta}{2}\left(\theta_i - \theta_{eq}\right)^2 + \sum_{torsions} \frac{V_n}{2}\left(1 + \cos\left(n\varphi - \varphi_0\right)\right) \tag{2}$$

$$V\left(r^N\right)_{Non-bonded} = \frac{q_i q_j}{D r_{ij}} + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right) \quad (3)$$

where r^N denotes geometry of the system, $V(r^N)$ is its potential energy, r_i and r_{eq} are the actual and equilibrium bond lengths, respectively, for the bond i, the θ_i and θ_{eq} is the same for bond angles, the φ and φ_0 is the same for dihedral angles, q_i and q_j are partial charges on the atom i and j, respectively; r_{ij} is distance of atoms i and j, D is dielectric constant and the remaining symbols are force field parameters.

Force field based scoring functions calculate the binding score as a sum of individual contributions made by various interactions in the bound complex. Force field based scoring functions commonly used in docking software mainly use non-bonded and torsion terms. The binding process normally takes place in water, so the desolvation energies of the ligand and the protein are sometimes taken into account implicitly. Since hydrogen bonding is one of the dominating interactions for the majority of complexes, some of the docking software, like AutoDock (Morris et al., 2009) and G-Score (Kramer et al., 1999), include a separate term for the treatment of hydrogen bonding.

B. Empirical scoring functions: Empirical based scoring functions, as in the case of force field methods, calculate the binding score of a complex as a sum of several weighted empirical energy terms that account for various types of non-bonded interactions. However, as opposed to the force field methods, empirical based scoring functions are much less systematic and general. The final score ΔG is calculated as a sum of weighted empirical energy terms, $\Delta G = \sum W_i * \Delta G_i$, where ΔG_i represents individual empirical energy terms, such as vdW energy, electrostatic energy, hydrogen bonding, desolvation, hydrophobicity, entropy etc., while W_i is the corresponding weight coefficient for a particular energy term, determined by linear fitting to an experimental data set. A set of X-ray receptor ligand complexes and their corresponding experimental binding energies are usually used as training data to calculate the weight coefficients by regression analysis. Due to the simple nature of the equation, these methods are computationally much more efficient compared to force field based methods. However, there are also significant drawbacks. General applicability of these functions is strongly dependent on the experimental data set used for their parametrization. It is not reliable to use such a scoring function for a data set that is structurally different from the training set. Glidescore (Halgren et al. 2004), LigScore (Krammer et al., 2005), and X-Score (Wang et al., 2002) are examples of software using empirical scoring functions.

C. Knowledge based scoring Functions: Knowledge based scoring functions use the sum of the potential of mean force (PMF) between the protein and the ligand, using data derived from 3D structure databases. These scoring functions are based on capturing the protein ligand atom pair frequency of occurrence in the structural database. It is assumed that each interaction type between a protein atom of type *i* and a ligand atom of type *j,* found at a certain distance r_{ij}, has an interaction free energy *A(r)*, which is defined by an inverse Boltzmann relation (Eq. 4).

$$A(r) = -K_b T \ln\left[\rho(r) / \rho^*(r)\right] \quad (4)$$

where K_b is the Boltzmann constant, T is the absolute temperature, $\rho(r)$ is the density of occurrence of the atom pair at distance r in the training set and $\rho^*(r)$ is this density in a reference state where the atomic interactions are zero.

The advantage of knowledge based scoring functions over empirical scoring functions is that there is no fitting to the experimental free energy of the complexes in the training set, whereas solvation and entropic effects are included implicitly. It should be noted that knowledge based scoring functions are used to reproduce the experimental structures rather than to predict binding energies. They can identify non-binders on their own or in combination with some other docking software during virtual screening. Since not all the possible interactions can be inferred from the crystal structure, these scoring functions may not be so robust and accurate, but they usually offer a good balance between speed and accuracy.

4.6 Techniques to improve the performance of scoring functions

Consensus scoring: This is a combination of the information obtained from different scores. The approach is helpful in balancing out the error of individual scoring functions, thus improving the probability of finding an appropriate solution. Several published studies show that combining the scores from different methods performs better than considering only the individual scores. MultiScore (Terp et al., 2001) and X-Score (Wang et al., 2002) are the most popular examples using consensus scoring.

Clustering: We often find an incorrect geometry with a slightly more favorable binding score than the correct geometry. However, these incorrect geometries are found with a very low frequency (~1-2%) when multiple docking experiments are performed. Thus, RMSD based clustering of all the docking solutions can be performed. To get the correct pose, the best energy conformation from the most populated cluster should be chosen.

4.7 Description of some commonly used docking programs

Table 1.1 summarizes the main features, license type and source for the most popular docking programs. We further provide a more detailed description of a few selected pieces of docking software. We would like to state that these methods are not necessarily the most accurate ones, but they are definitely the most widely used and the most cited in the docking community.

AutoDock: AutoDock3 (Morris et al., 1998) and AutoDock4 (Morris et al., 2009) are force field based docking programs which have been widely used for the automated docking of small molecules, such as peptides, enzyme inhibitors and other ligands, into macromolecules, such as proteins, enzymes and nucleic acids. AutoDock offers optimization procedures like simulated annealing, genetic algorithm (GA) for global searching, a local search (LS) method to perform energy minimization, or a combination of both (GALS) for getting the accurate docked complex. The scoring function used in AutoDock is inspired by the MD programs AMBER, CHARMM or GROMOS, it includes terms for the Lennard-Jones potential, Coulombic electrostatic potential, hydrogen bonding, partial entropic contribution, desolvation upon binding and a hydrophobic effect. The scaling parameters for these terms were derived from a set of 30 protein-ligand complexes.

Software	Ligand sampling methods[a]	Receptor sampling methods[a,b]	Scoring function[c]	Solvation scoring[b,d]	License type[e]	Source
AutoDock3	SA, GA	NA	MM+ED	DDS, DS	FAS	(Morris et al., 1998)
AutoDock4	SA, GA	SE	MM+ED	DDS, DS	FAS	(Morris et al., 2009)
AutoDock Vina	CB	CB	ML	NA	OPS	(Trott et al., 2009)
DOCK6	IC	SE	MM	DDD/GB/PB	FAS	(Kuntz et al., 1982)
ICM	MC	MC	MM+KB	DDD,PBE,DS	CPL	(Abagyan et al., 1994)
Glide	CE+MC	TOS	MM+ED	DS	CPL	(Halgren et al., 2004)
GOLD	GA	NA	MM+ED	NA	CPL	(Jones et al., 1995, 1997)
FlexX/FlexE	IC	SE	MM+ES	NA	CPL	(Rarey et al., 1996)

[a]Sampling methods can be Genetic Algorithm (GA), Conformational Expansion (CE), Monte Carlo (MC), Simulated Annealing (SA), Molecular Dynamics (MD), Incremental Construction (IC), Merged Target Structure Ensemble (SE), a combination of GA, SA and MC (CB), and Torsional Search (TOS); see Section 4.4 for more information. [b]If the package does not accommodate this option, the symbol NA (not available) is used. [c] Scoring functions can be Empirical (ES), Knowledge Based (KB) or force field (MM) based; see Section 4.5 for more information. [d]The accuracy of the scoring function can be improved using implicit solvent models. Solvation scoring can be done using Distance-Dependent Dielectric (DDD), Poisson Boltzmann Dielectric (PBE), a parameterized desolvation term (DS), Generalized Born (GB), and linearized Poisson Boltzmann (PB) equations. The license type can be [e]Freely available (FAL), Open Source (OPS) and Commercial Paid License (CPL) for academic users only.

Table 1. Details of Commonly Used Docking Software.

The advantage of AutoDock4 over AutoDock3 is that it allows receptor flexibility, and also an improved new force field is used to calculate the binding energy. The force field of AutoDock4 includes a new intramolecular term, and a full desolvation model for desolvating polar and charged atoms. AutoDock facilitates the clustering of all the docked orientations by defining a root mean square tolerance, which can also be used to find the potential binding regions. It was seen that the lowest energy structure in the most populated cluster successfully reproduces the crystal structure.

AutoDock Vina (Trott & Olson, 2009): It is a new generation of docking software (referred to as Vina) from the Molecular Graphics Lab, the developer of the other versions of AutoDock. It is a user friendly, open source piece of software, capable of predicting binding modes with better accuracy, while it is significantly faster than AutoDock4. It uses a combination of optimization algorithms, such as the genetic algorithm, swarm optimization and simulated annealing, to place the ligand in the binding site. The scoring function used in Vina is more based on machine learning rather than directly on a force field. Similarly to AutoDock4, it allows receptor flexibility.

The philosophy behind the development of Vina was to make the software easy to use, so most of the parameters used during docking are set by default, reducing the possibility of making manual mistakes. A further speed up in docking is achieved by multithreading. Thus, overall, Vina is very suitable for docking a large set of different compounds.

DOCK: The program package DOCK (Kuntz et al., 1982, currently version 6.4) basically works in a few subsequent steps. First, the program "sphgen" is employed in order to identify a binding site and to generate spheres within the active site. Secondly, the program "grid" is used to generate scoring grids. Then the last program "DOCK" matches the sphere with the ligand atoms and uses the scoring grid to evaluate the ligand orientation. It constructs the ligand in the binding site step by step using the Anchor-and-Grow algorithm. Initially, the rigid anchor fragment of the receptor is placed at a selected position, and then is gradually enlarged by adding the flexible fragments of the ligand. An additional extension to DOCK allows rescoring the docked configuration of the ligand using several secondary scoring functions.

ICM: The Internal Coordinates Mechanics (ICM) software is a set of modules for various purposes, such as visualization, chemical drawing and editing, homology modeling, docking and virtual screening (Abagyan et al., 1994). The ICM-Docking and chemistry module performs flexible ligand docking in a grid based receptor field. The scoring function used in ICM primarily accounts for electrostatics, van der Waals, hydrogen bonds, and the hydrophobic term. ICM needs the protein structure to be converted into an ICM object before docking. It provides a simple, object based GUI which can be used for docking. The binding site can be defined by entering the binding site residues, using the graphical selection tool, or the implemented icmPocketFinder function. It generates receptor maps within the defined boundary, which are further used in the docking. It is necessary to define the initial position where sampling will begin. ICM facilitates interactive, as well as batch docking. In interactive docking, one ligand is docked at a time in the foreground, whereas batch ligand docking runs in the background and is thus ideal for large scale docking jobs and virtual screening of huge ligand libraries. ICM offers an attractive feature to visualize and browse the docking results, and scan the hit compounds.

BALLDock/SLICK: BALLDock/SLICK (Kerzmann et al., 2008) is specially designed for docking of carbohydrate like compounds, with applications in carbohydrate based drug design. Molecular docking of protein-carbohydrate complexes needs some special attention because of the special features of such interactions, such as the unusual flexibility of carbohydrates, stacking interactions with aromatic amino acids and a high number of hydrogen bonds involved in binding. Protein carbohydrate interactions are strongly influenced by CH$\cdots\pi$ interactions, which are mostly ignored in the commonly available scoring functions and are considered in BALLDock/SLICK. This docking program uses genetic algorithms to search the configurations of the ligand within the defined search space, and the scoring function SLICK is used to calculate the binding score of the docked conformations. Kerzmann et al. compared the performance of BALLDock with FlexX on a set of 22 lectins and sugar-binding proteins complexed with carbohydrate. FlexX achieved good results but still did not reach the predictive accuracy of BALLDock/SLICK.

TRITON: We previously introduced our in house graphical tool TRITON, and mentioned its use for homology modeling and mutagenesis. Another functionality of TRITON relates to *in*

silico engineering of protein-ligand binding properties (Prokop et al., 2008). The program can be used as a graphical user interface for the docking software AutoDock3, AutoDock4 and Vina. It enables the user to do common pre-docking tasks, like creating a project directory, reading structures, manipulating structures, calculating various types of charges and finally preparing input files for docking. Docking wizards make the job easy for new users, where the step by step procedure decreases the possibility of missing any of the docking parameters. It includes and offers certain optimized docking parameters that can be used as basic starting points if the user is not sure about certain docking parameters used in the AutoDock suite of programs. TRITON also includes parameters for ions taken from case specific studies, so it is easy to handle ions during the docking. Another important feature of TRITON is that it facilitates interactive analysis and visualization of the docking results.

5. Free energy calculation

As discussed above, various docking software can successfully predict the correct binding mode of the ligand into the receptor (Taylor et al., 2002; Warren et al., 2006). However, the previously described empirical scoring functions, which are based on a single receptor/ligand structure, do not provide accurate enough predictions of the binding free energy (ΔG), the key quantity characterizing the strength of the receptor/ligand interaction. To tackle this problem, molecular dynamics (MD) or Monte Carlo (MC) based methods for free energy calculation were developed in the mid 1980s (Jorgensen & Ravimohan, 1985). These methods, formally rooted in statistical thermodynamics, are now frequently used to compute receptor/ligand binding free energy. The methods use molecular mechanics force fields and Newtonian physics to evaluate the dynamics of the system. In the case of MD, we follow the evolution of the dynamics of the system in time. The dynamics allows the system to accommodate various protein side chains as well as ligand conformations, and also ligand configurations with respect to the protein. Simulations are usually performed in the bound state. Here, we will discuss the methods most commonly used to evaluate the binding free energy between the receptor and the ligand, namely Free Energy Perturbation (FEP), Thermodynamic Integration (TI), and Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA). We will also give some notes about the combined molecular mechanics/quantum mechanics (QM/MM) techniques.

5.1 Free Energy Perturbation (FEP) and Thermodynamic Integration (TI)

The FEP and TI approaches for free energy calculation are based on statistical thermodynamics and are generally formulated not to calculate the absolute value of the free energy, but always a relative value, i.e., the free energy difference, ΔG, between two equilibrium states. This is of a great importance, since for *in silico* mutagenesis applications we always need only relative values.

The FEP and TI free energy calculations are carried out using a thermodynamic cycle. Such a cycle, adapted for *in silico* mutagenesis purposes, is shown in Fig. 6. It involves a mutation of either the receptor alone, or the receptor/ligand complex (start state) into another state (end state), where the receptor is mutated. The simulation can be performed in either implicit or explicit solvent. The final calculated quantity is $\Delta\Delta G$. This number will tell us whether the mutated protein (P_M) exhibits higher or lower affinity to the ligand L compared to the wild type protein (P_W).

As the start and the end state can be arbitrarily different, these calculations are sometimes referred to as *computational alchemy.*

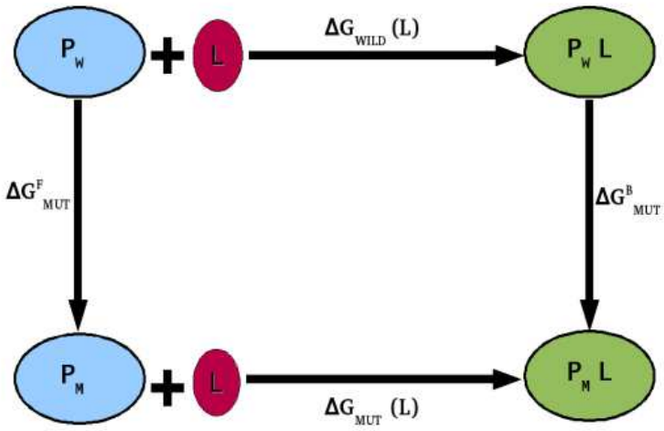


Fig. 6. Thermodynamic cycle for calculating the relative binding free energies of a ligand L to mutated system (P_M). ΔG^{F}_{mut} is free energy change between the wild type and mutated receptor, ΔG^{B}_{mut} is free energy change between the wild type receptor/ligand complex and the mutated receptor/ligand complex, $\Delta G_{wild}(L)$ and $\Delta G_{mut}(L)$ are binding free energies for the wild type receptor/ligand and mutated receptor/ligand complexes, respectively.

As the free energy is the state function, Eq's (5 and 6) must hold.

$$\Delta G_{wild}(L) + \Delta G^{B}_{mut} = \Delta G^{F}_{mut} + \Delta G_{mut}(L) \tag{5}$$

$$\Delta\Delta G = \Delta G_{mut}(L) - \Delta G_{wild}(L) = \Delta G^{B}_{mut} + \Delta G^{F}_{mut}(L) \tag{6}$$

The **FEP** calculations are based on the Zwanzig's formula (Zwanzig, 1954) to calculate the free energy difference ΔG between two states (see Eq. 7).

$$\Delta G^{FEP} = G_B - G_A = -k_B T \ln\left\langle \exp\left(\frac{V_B - V_A}{k_B T}\right)\right\rangle_A \tag{7}$$

where k_B is the Boltzmann constant, T is the absolute temperature, $\langle\rangle_A$ denotes the MD or MC ensemble average over a simulation run for state A, V_A and V_B are the potential energies of state A and B, respectively. In general, an ensemble is an average set of systems, that are identical in all respect apart from the dynamics of the atom (k/a ensemble), considered all at once, each of which represents a possible state that the real system might be in. The potential energy difference can be averaged over an ensemble generated using the start and end state potential function for the forward and backward process, respectively.

The goal is to obtain the convergence of the values resulted from Eq. 7 within a reasonable time. It is assumed that the relevant geometries sampled on the potential energy of state A have a considerable overlap with those of state B.

The transition of state A into state B may also yield high energy geometries in the complex because of steric clashes with the neighbouring atoms. To overcome this issue, transition is done via many non-physical intermediate states that are usually constructed as a linear combination of the potential calculated for the start and end state. The potential energy of an intermediate state between A and B is given as shown in Eq. 8,

$$V_\lambda = (1-\lambda)V_A + \lambda V_B \tag{8}$$

where λ varies from 0 to 1. This state is a hypothetical mixture of states A and B: when λ=0, $V_\lambda = V_A$, and when λ=1, $V_\lambda = V_B$. Therefore, the transformation of state A into state B is done smoothly, by changing the values of the parameter λ in small increments, dλ. In practice, the free energy difference between the states A and B is computed by summing over all the intermediate states along the λ variable (Eq. 9).

$$\Delta G = \sum_i dV\lambda i \tag{9}$$

This approach of breaking down the transitions into multiple smaller steps shares similarity with another approach used to compute free energy, namely Thermodynamic Integration (TI) (Kirkwood, 1935). **TI** is based on integrating a different equation from statistical thermodynamics, where the free energy difference between two states is obtained by integrating the derivative of the mixed potential function over λ (Eq. 10).

$$\Delta G^{TI} = \int_0^1 \left\langle \frac{\delta V(\lambda)}{\delta\lambda} \right\rangle_\lambda d\lambda \approx \sum_i w_i \left\langle \frac{\delta V(\lambda)}{\delta\lambda} \right\rangle_{\lambda_i} \tag{10}$$

In this case, the mixed potential V(λ) is defined numerically by evaluating the linear interpolation between the potential function of the start and end state, respectively.

In principle, both FEP and TI should give the same results, as the free energy is a state function.

The relative binding free energy difference ΔΔG between the wild type protein P_W and its mutant P_M can easily be calculated from Eq's. 5 and 6 (for denotation see Fig. 6), and where ΔG^F_{mut} and ΔG^B_{mut} are calculated using the above described FEP or TI methods for the free and bound state, respectively.

As mentioned before, with the FEP or TI approach, the free energy associated with the two unphysical paths $P_W \rightarrow P_M$ (mutation in the free state) and P_W (L) → P_M (L) (mutation in the bound state) is calculated by sampling the degrees of freedom of the free protein or the complex using molecular dynamics (MD) or Monte Carlo (MC) methods. At regular intervals, the atoms of the residue which is being mutated are replaced by atoms of the residue which is desired at that place, and the potential energy along the paths is recorded. This quantity, averaged over the complete simulation, gives a proper free energy change ΔG_{mut}. However, the convergence of the free energies is a first critical issue in the accurate calculation of the binding free energy. This requires exhaustive sampling of the system, which is much more time consuming than docking or normal MD simulations. Moreover, the mutation may cause steric clashes with the neighbouring atoms, which makes the sampling issue even more complicated.

5.2 Molecular mechanics poisson-boltzmann surface area (MM-PBSA)

Another approach well suited for estimating the binding free energy of molecular complexes and their mutants is the Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) method (Srinivasan et al., 1998). The MM-PBSA approach was initially used to study the stability of nucleotide fragments, but also to compute the relative or absolute binding free energy of protein-ligand complexes. Later extensions (see Kollman et al., 2000; Hou et al., 2011) have enabled employing the method for free energy calculation in *in silico* mutagenesis approaches, which is helpful in making predictions for protein engineering. Unlike FEP and TI, MM-PBSA is an endpoint method that calculates binding free energy without consideration of any intermediate state.

The MM-PBSA approach is used to calculate the free energy change ΔG_{bind} upon ligand binding according to equation 11. It combines the molecular mechanical energies with the continuum solvent approaches, and approximates the average of each state in order to calculate the binding free energy.

$$\Delta G_{bind} = < G_{complex} > - \left(< G_{receptor} > + < G_{ligand} > \right) \tag{11}$$

The single terms are defined by Eq's 12-14.

$$G_X = H - TS = E_{MM} + G_{sol} - TS \tag{12}$$

$$E_{MM} = E_{internal} + E_{electrostatic} + E_{vdw} \tag{13}$$

$$G_{sol} = G_{PB/GB} + G_{SA} \tag{14}$$

where X stands for the complex, receptor or ligand, T is the absolute temperature. The E_{MM}, G_{sol} and S are the gas phase molecular mechanics energy, solvation free energy and entropy, respectively. The E_{MM} includes several energy terms: $E_{Internal}$ for bond, angle and dihedral contributions, $E_{electrostatics}$ for coulomb interactions and E_{vdw} for van der Waals energies. G_{sol} is the sum of electrostatic solvation energy and non-polar contribution to the solvation free energy. The electrostatic contribution to the solvation free energy is calculated by solving either the linearized Poisson Boltzman (PB) or Generalized Born (GB) equation, while the non-polar contribution is estimated from the solvent accessible surface area (Connolly, 1983). If the solvation free energies are computed from the Generalized Born (GB) model, the method is termed also MM-GBSA. The last term, TS, includes the solute entropy S, which is usually calculated by quasi-harmonic analysis of the snapshots using normal mode analysis (Srinivasan et al., 1998).

Ideally, this approach is based on post-processing molecular dynamics trajectories. The free energy contributions are calculated for each component of the system (protein, ligand and the complex) from the snapshots taken from MD trajectories. In order to get the binding free energy of a ligand, two alternatives are used (see Fig. 7). The first is a multi trajectory approach, where we use the trajectories from three separate molecular dynamics simulations (on the complex, receptor and ligand). Snapshots of each component (protein, ligand and complex), taken from their corresponding simulation trajectories, are used to calculate the free energy terms. Note that this approach takes into account the influence of conformational changes upon binding on the final binding free energy.

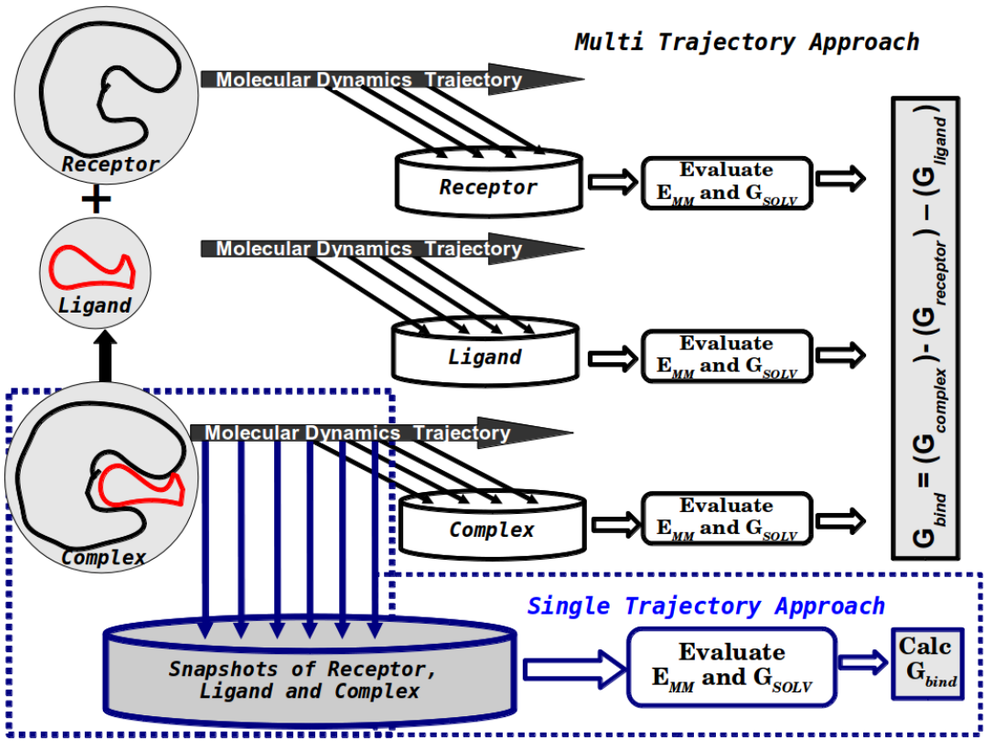


Fig. 7. Diagram for MM-PBSA and MM-GBSA calculations on a solvated complex. Single trajectory approach is surrounded by a blue dotted line.

In the second approach, molecular dynamics simulations are run on the complex only, in order to reduce noise and cancel out the errors in the simulations. Conformational snapshots for the receptor alone and the ligand alone are extracted from the MD simulation of the complex by removing the respective binding partner from the complex. Therefore, it is assumed that the structure of the receptor and ligand is the same in the bound and the free state, and no major conformational changes occur upon binding. In this approach, $E_{Internal}$ is canceled out between the complex, protein and ligand, which reduces the noise in calculations.

In principle, the first approach of running three independent molecular dynamics simulations of three species is more accurate than the single trajectory approach. In practice though, the multi trajectory approach seems not to be used extensively. This is understandable, since there is no proper way to get the convergence of E_{MM} values for the receptor within reasonable computational time. Hence, the regular implementation of this method is usually based on the second approach, where only the MD trajectory of the complex is used to compute the binding free energy.

A fundamental issues associated with the MM-PBSA approach is entropy calculation. The normal mode analysis (NMA) approach is usually employed to calculate entropy. However, this approach overestimates the loss of entropy upon ligand binding. In order to get meaningful absolute binding free energies, the entropy contribution must be determined in a consistent fashion. The best approach is to compute the relative binding free energies of a series of similarly sized ligands, where the entropy contribution is expected to cancel out

(Massova & Kollman, 1999). The *in silico* mutants of a protein are also expected not to have a significant change in entropic contribution to the binding.

5.3 Alanine scanning mutagenesis using MM-PBSA

In the *in silico* mutagenesis section we discussed the basis of the methodology of performing single point or multiple alanine mutations using computational approaches. Here we will only show how to use alanine scanning mutagenesis coupled with the MM-PBSA approach. We are distinguishing between two complementary problems of mutagenesis where binding free energy is calculated by the MM-PBSA approach. The first refers to the change in binding free energy upon alanine mutation at any location. This can be solved using the previously described single trajectory MM-PBSA approach on two systems, the wild type and the mutant. Molecular dynamics simulations of two systems (ligand complexed with the wild type and with the alanine mutant, respectively) are run under the same conditions. These two different trajectories are subjected to the MM-PBSA calculation previously described. The change in binding free energy upon mutation is now the difference between the binding free energy of the mutant and that of the wild type. In principle, this approach is accurate and recommended because it samples the conformational changes of the system upon mutation and takes into account their effect on the change in free energy.

The second issue refers to the individual contribution of each residue to the binding. MM-PBSA was first used in this respect in a study where a single MD simulation was used to compute the individual contribution of each residue to the binding in protein−protein complexes. Snapshots of mutants are generated from a single molecular dynamics trajectory of a wild type system. Mutations are performed by removing side chain atoms beyond the β carbon of the amino acids under investigation (Massova & Kollman 1999). These snapshots are used for binding free energy calculations by the MM-PBSA approach. The approach used in alanine scanning mutagenesis is depicted in Figure 8.

On the one hand, this is not a very accurate way to get the change in free energy upon alanine mutation. On the other hand, it is very fast, as the mutations can be performed at any location without running the molecular dynamics simulation of the mutant system. Therefore, once we have the MD trajectory of the wild type, a possible primary scan for all the locations could be done in minimal computational time. Since the method uses the MD trajectory of the wild type to create the mutants, it is assumed that the receptor/ligand complex adopts the same geometry upon the mutation. This is a limiting factor, as mutations of the residues around the ligand binding site can substantially affect the binding geometry. Nevertheless, it is expected that this approach can estimate the free energy contribution made by a particular residue compared to the wild type system (Moreira et al., 2007). This approach is mainly recommended for finding the hot spots in protein-protein interactions. Hot spots are residues which make a contribution of about 2.0 Kcal/mol to the total binding free energy of the system, and are very important from the protein engineering point of view because they can be used as key points to alter the protein's recognition ability. Alanine scanning also gives an idea about the residues which are close to the binding region, but do not contribute substantially to the binding energy. These locations in the protein can be used to make the binding stronger if a residue with favourable properties is placed at that location. The MM-PBSA approach is quite fast. The calculations for hundreds of ligands and hundreds of mutants are feasible using high

performance computing facilities. One must mention here that these approaches are approximate, and the relevant predictions should be verified using FEP or TI before experimental trials.

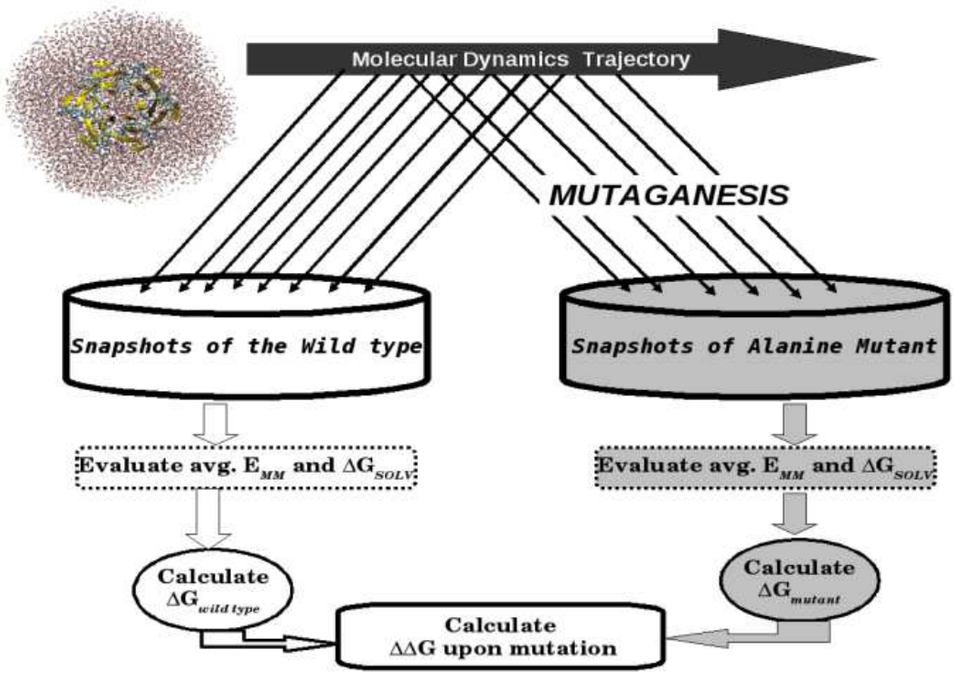


Fig. 8. showing a single trajectory alanine scanning mutagenesis approach used with MM-PBSA or MM-GBSA calculations.

5.4 Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) approaches

A combination of quantum mechanics and molecular mechanics (QM/MM), accompanied by the increasing computational power of modern parallel and vector-parallel platforms, has brought a real breakthrough in the simulation of large systems. Here we describe the current QM/MM strategies used for quantifying the binding energy of complexes involved in molecular recognition.

The seminal contribution made by Warshel et al. in 1976 marks the beginning of the QM/MM era (Warshel & Levitt, 1976). In brief, to model large biomolecules one uses a QM method to model the active region (originally substrates and co-factors of an enzymatic reaction), and an MM method for the treatment of the surroundings (e.g., protein and solvent). The QM/MM approaches are relatively new in the field of molecular docking. A few years ago, a combined QM/MM docking approach for the investigation of protein ligand complexes was presented for the first time, and very promising results were obtained by combining the fast docking technique with the subsequent QM/MM optimization of the docked structure (Beierlein & Clark, 2003). Later, in an attempt to develop a docking algorithm which can predict poses accurately for the cases where the conventional approach fails, QM/MM calculations were integrated in the scoring phase (Cho et al., 2005). A protein-ligand docking study of 40 complexes investigated through QM/MM based docking calculations suggests that the use of fixed charges during the docking exhibits on-trivial

errors. Therefore, polarization of the QM region is suggested to be crucial for docking studies. It was found that including also some protein atoms in the QM region, along with the ligand atoms, increases the success rate of QM/MM docking procedures (Cho & Rinaldo, 2009).

There are also examples in literature where a QM/MM approach was used to calculate the binding free energy. For example, Gräter and coworkers evaluated the performance of a QM/MM approach combined with MM-PBSA to obtain the protein/ligand binding free energy for a set of 47 benzamidine derivatives binding to trypsin. The QM/MM-PBSA methods reproduced the experimental binding energy well, with a root-mean-square (RMS) error of 1.2 kcal/mol (Gräter et al., 2005). Later, QM charge densities were used to solve the PB equation in a test case of binding of balanol and its derivative to the protein linase A (Wang & Wong, 2007). Even if this approach is not being used very frequently in the field of binding free energy calculation, the availability of packages (e.g. Amber Tools 1.5) that facilitate such a QM/MM-PBSA calculation in protein/ligand complexes, along with recent developments, is expected to make the QM/MM-PBSA method more user friendly.

Nowadays, all the statistical mechanics techniques to determine free energy differences through sampling, e.g., TI, umbrella sampling, or FEP are being used in conjunction with semiempirical QM/MM methods (Chung et al., 2009; Tuttle, 2010). The continuous increase in computer power has played an essential role in the development of these methods. QM/MM methods are expected to be especially important in the field of molecular recognition for systems where ions are present, i.e., in the area of metalloproteins.

6. Case studies

We present a few studies where *in silico* protein engineering was successfully used to study molecular recognition. We compare our results with other recently published studies on altering the binding specificity of a receptor by using *in silico* mutagenesis. The citations particular to the studied systems and to the methods mentioned below are omitted here, and can be found in the respective papers.

6.1 Engineering of the PA-IIL lectin to understand its sugar preference

Lectins are proteins of non-immune origin that recognize carbohydrates with high specificity and affinity. They belong to a large family of proteins whose unifying feature is the ability to decode the information stored in the glycome. Lectins are involved in a diverse set of biological processes, such as cell-cell recognition, differentiation, signalling or the adhesion of infectious agents to host cells. Many of these functions are connected with the recognition of specific saccharide structures on the cell surface. Carbohydrate-binding proteins play a key role also in host cell recognition by pathogens, as their specific adhesion to the host cell tissue is the first stage of their infectivity. Thus, lectins from pathogens represent a primary target for anti-adhesion therapy, having a great potential in the field of drug design.

The selected study (Adam et al. 2007, 2008) is based on the *in silico* protein engineering of the protein PA-IIL, a lectin from an opportunistic human pathogenic bacterium *Pseudomonas*

aeruginosa, which causes lethal complications in cystic fibrosis patients. PA-IIL is a tetrameric lectin characterized by an unusually high (micromolar) affinity to L-fucose, which is atypical in protein-carbohydrate binding. Lectins homologous to PA-IIL later identified in other microorganisms, display high sequence and structure similarity, but strongly differ from each other in terms of sugar preference. For example, the lectin RS-IIL from *Ralstonia solanacearum* strongly prefers D-mannose over L-fucose. Three amino acid residues, at positions 22−23−24, were identified as the key residues that describe the relationship between structure and binding specificity for these lectins, and were named the "specificity binding loop". Given the capital relevance of this loop, *in silico* approaches were applied to understand the precise role of the specificity loop in the sugar binding preference.

PA-IIL				
saccharide	Expt[a]	AD3[b]	DOCK (Std)[b]	DOCK (Amber)[b]
Me-α-L-Fuc	−8.71	−10.7	−40.01	−32.22
Me-α-L-Gal	−8.09	−9.93	−40.02	−28.92
α-L-Fuc	−6.94	−9.33	−35.92	−27.56
Me-α-D-Man	−5.97	−9.23	−43.50	−26.88
S22A				
Me-α-D-Man	−7.58	−10.47	−32.60	−22.30
Me-α-L-Fuc	−7.39	−9.26	−30.60	−23.47
α-L-Fuc	−6.84	−8.96	−28.81	−19.85
Me-α-L-Gal	−6.60	−9.49	−31.70	−20.34
S23A				
Me-α-L-Fuc	−9.04	−10.49	−37.90	−20.82
Me-α-L-Gal	−8.11	−10.79	−38.83	−19.86
α-L-Fuc	−7.32	−8.98	−35.43	−18.81
Me-α-D-Man	−5.84	−9.15	−39.95	−16.80
G24N				
Me-α-L-Fuc	−9.19	−9.65	−39.84	−28.19
Me-α-L-Gal	−8.06	−9.83	−38.42	−22.26
α-L-Fuc	−7.18	−9.29	−37.10	−22.48
Me-α-D-Man	−5.96	−9.15	−45.03	−25.16

Table 2. Experimental (Expt) and calculated energies of monosaccharides binding to PA-IIL and its mutants obtained from AutoDock3 and DOCK (all values in Kcal/mol). AD3 stands for AutoDoc3 binding energies, DOCK (std) for energies from inbuilt evaluation of the DOCK software and DOCK (Amber) for energies from DOCK reevaluated by AMBER. Saccharides used: α-L-Fuc (α-L-Fucopyranose), Me-α-L-Fuc (Me-α-L-fucopyranoside), Me-α-L-Gal (Me-α-L-galactopuranoside), Me-α-D-Man (Me-α-D-mannopyranoside). Values taken from [a]Adam et al, 2007 and [b]Adam et al, 2008

The dimeric structure of PA-IIL was used as a template structure for the homology modeling of three single-point mutants (S22A, S23A, and G24N matching amino acids in RS-IIL) of PA-IIL using our *in house* developed software TRITON, interfaced with MODELLER. In order to understand the role of a particular mutation with respect to sugar preference, different monosaccharides were docked into PA-IIL and its mutants using AutoDock3 and DOCK. Since PA-IIL has two Ca^{++} ions in the binding site, which mediate the sugar binding, the effect of their charge on the docking energy was also evaluated. A formal charge on Ca^{++} equal to 1.8 and 2.0 gave results in good agreement with experiment. The value of 1.8 was chosen as a compromise, because Ca^{++} surrounded by several negatively charged oxygen atoms adopts a smaller charge in reality (about 1.5, see Mitchell et al, 2005).

The docking simulations produced a series of binding energies for the possible complexes of saccharides bound to PA-IIL and its mutants. The results can be seen in Table 1.

Overall, the docking results from AutoDock3 confirm that PA-IIL has higher preference for Me-α-L-Fuc (-10.7 Kcal/mol) over Me-α-D-Man (−9.23 Kcal/mol), and the sugar preference switches from Me-α-L-Fuc (-9.26 Kcal/mol) to Me-α-D-Man (-10.47 Kcal/mol) upon the S22A mutation. Docking inside two other mutants S23A and G24N also shows the order of preference again similar to what can be observed experimentally. Qualitatively, DOCK overestimates the binding energy in more cases than AudoDock3. Compared to experimental results, the AutoDock3 results reproduced the experimental order of saccharide preference to a large extent. The authors conclude that the automated docking methods are capable of identifying preference trends, and, therefore, using *in silico* approaches in pre-planning the *in vitro* mutations can help to identify the best potential candidates for mutagenesis.

6.2 Double mutant avian H5N1 virus hemagglutinin

The study (Das et al., 2009) shows how the free energy perturbation approach is used to compute the binding affinity of hemagglutinin (HA) to sialylated glycan epitops. A typical influenza infection, caused by avian influenza A viruses (H1N1, H2N2, H3N2 and H5N1 subtypes), requires binding of the viral surface glycoprotein hemagglutinin (HA) to sialylated glycans present on the host cell surface in order to initiate the infection. A change in the binding specificity of the HAs from α-2,3 (common in avians) to α-2,6-linked (common in human) sialylated glycans is expected to facilitate transmission of the virus from avians to humans. Therefore, molecular recognition of the particular glycans, considered as a key point for such infections, was inspected using mutagenesis studies.

HAs are homo trimers, with each monomer comprising of two subunits. The Receptor Binding Domain (RBD) of HAs, formed by basically 3 loops, requires at minimum two mutations to switch receptor specificity from avian to human. It is also known that hemagglutinin H1 changes its specificity from human to avian epitopes after two mutations (D190E and D225E). The authors were interested in finding whether a double mutation in hemagglutinin H5 enable it to recognize the human analog, as it is seen for the H1 HA subtype.

The authors used *in silico* approaches to interpret and predict the critical mutations responsible for HA-receptor binding. In order to achieve this, the change in relative binding affinity of H5 HA to sialylated glycans upon mutation was calculated through free energy

perturbation approaches. The change in binding energy due to a mutation is evaluated using a thermodynamic cycle (see Fig. 6), where $\Delta\Delta G_{bind}$ is calculated from the free energy change caused by the same mutation for the bound and free states respectively. This simulation was performed over 22 λ points, where each window was simulated for 0.3 ns. Therefore, a total of 66-ns of simulation were performed for each mutation in order to get proper sampling. The authors claim that before analyzing the effect of novel mutations on the H5 HA receptor, they validated their protocol by comparing the calculated binding affinities against experimental data for other mutants.

The authors conclude that the FEP calculations are in a fairly good agreement with the glycan array data, which was available for only a few H5 HA mutants. Most of the evaluated mutations resulted either in no change, or in weak binding affinity to α-2,6-linked sialylated glycans compared to α-2,3. They identified that a double mutation (V135S and A138S) in H5 HA enhances the specificity towards α-2,6-linked sialylated glycans: $\Delta\Delta G_{bind}$= -2.56+-0.73 Kcal/mol for the human receptor, compared to $\Delta\Delta G_{bind}$= 0.84+-1.02 Kcal/mol for the avian receptor. To validate the results, the authors repeated the calculations for the same mutants on H5 HA obtained from a different isolate, which also revealed a substantial increase in the binding affinity for the human receptor. In order to understand the forces behind the recognition, they performed a free energy component analysis and saw that the electrostatic interactions are the driving forces for change in binding specificity upon mutation.

Thus, this study used computational approaches to provide valuable insight into the molecular recognition of glycans. This is another example where *in silico* protein engineering approaches were used as a complementary tool to interpret and understand molecular recognition.

6.3 Structural basis of NR2B-selective antagonist recognition

The third example (Mony et al., 2009) gives a detailed characterization of the ifenprodil binding site in the NMDA receptor (NMDAR) by both *in silico* and *in vitro* approaches. The NMDA receptor is an ionotropic glutamate receptor, which serves as the predominant molecular device for controlling synaptic plasticity and memory function. Therefore, controlled activation of the NMDA receptor is of great interest as a potential therapeutic target.

In order to stop receptor overactivation, several NMDAR competitive antagonists were developed in the 1980s. However, these compounds failed in clinical trials because of their inability to discriminate between the various NMDAR subtypes, and caused a generalized inhibition. In the study we report here, the authors used the most promising NMDAR antagonist at that time, ifenprodil, and its derivatives, in order to characterize the ifenprodil binding site using both computational and experimental approaches. The ifenprodil binding site on NMDAR was mapped on NR2B subunit's N-terminal domain (NR2B NTD), and the authors were able to describe the structural determinants responsible for the high-affinity binding of ifenprodil on the NR2B subunit.

A homology modeled structure of NR2B NTD was generated using the sequence to structure alignment functionality within the comparative modeling tool of MODELER 9.0. The ifenprodil was docked into the modeled structure using LigandFit. During the docking, the structure of the protein was kept rigid and 20 conformers of the ligand were subjected to

energy minimization in the molecular modeling tool CHARMM. A 1 ns MD simulation of the minimized structures was used to generate the pharmacophore model of the system. In this case the *in silico* approach was used to get a clear picture of the system before extensive experimental validation was achieved by site directed mutagenesis.

Docking showed that ifenprodil adopts a unique and well defined orientation in the central crevice of the NR2B NTD. Based on the *in silico* model, site directed mutagenesis proved 5 NR2B NTD residues (Thr76, Asp77, Asp206, Tyr231, Val262) are essential for the high affinity ifenprodil binding and receptor inhibition. The proposed model of ifenprodil binding to NR2B NTD shared some similarities with a previously proposed model, which had had no experimental validation (Mirienelli et al., 2007). The authors suggest that the differences in the models could be caused by the use of different sequence alignment for the loops situated in binding cleft. However this study showed that a suitable combination of *in silico* approaches can provide a good picture of what we can expect before starting any kind of experiment.

7. Concluding remarks

We have shown in this chapter how *in silico* protein engineering can be used in the field of molecular recognition. The particular steps one has to go through when using these techniques were described. They comprise of 3D structure determination, *in silico* mutagenesis, docking as the first approximation of the binding affinity, and, finally, accurate calculation of the binding free energy.

It should be highlighted that, in many cases, *in silico* approaches provide information complementary to that obtained by experimental approaches. A number of such methods have been implemented and are available in specialized software packages. Therefore users can test the different tools easily and select the ones able to perform well for the particular system they are interested in. We have provided also a brief list of the most frequently used computer programs for the particular tasks described. It is probably fair to say that *in silico* approaches are mostly useful for the visualization and intelligent design of protein engineering projects. As the computer power increases and software products become more and more sophisticated, it is highly probable that *in silico* protein engineering of proteins recognizing small molecules will become an even more useful tool in the future.

8. Acknowledgment

The authors thank the Ministry of Education of the Czech Republic (Contracts MSM0021622413, ME08008) and Czech Science Foundation (Contracts 303/09/1168, P207/10/0321, 301/09/H004) for financial support. This work was further supported by the project "CEITEC - Central European Institute of Technology" (CZ.1.05/1.1.00/02.0068) from European Regional Development Fund.

9. References

Abagyan, RA. & Batalov, S. (1997). Do aligned sequences share the same fold ?. *Journal of Molecular Biology*, Vol. 273, No. 1, pp. 355-368.

Abagyan, R.; Totrov, M. & Kuznetsov, D. (1994). ICM-A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, Vol. 15, No. 5, pp. 488-506.

Adam, J.; Pokorná, M.; Sabin, C,; Mitchell, EP.; Imberty, A. & Wimmerová, M. (2007). Engineering of PA-IIL lectin from pseudomonas aeruginosa - unravelling the role of the specificity loop for sugar preference. *BMC Structural Biology*, Vol. 7:36.

Adam, J.; Kríz, Z.; Prokop, M.; Wimmerová, M & Koca, J. (2008). In silico mutagenesis and docking studies of *Pseudomonas aeruginosa* PA-IIL lectin predicting binding modes and energies. *Journal of Chemical Information and Modeling,* Vol. 48, No. 11, pp. 2234-2242.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389-3402

An, J.; Totrov, M. & Abagyan, R. (2005). Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & Cellular Proteomics: MCP*, Vol. 4, No. 6, pp. 752-761.

Baxter, CA.; Murray, CW.; Clark, DE.;Westhead, DR. & Eldridge, MD. (1998). Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins*, Vol. 33, No. 3, pp. 367-382.

Beierlein, F.; Lanig, H.; Schurer, G.; Horn, AHC. & Clark T. (2003). Quantum mechanical/molecular mechanical (QM/MM) docking: an evaluation for known test systems. *Molecular Physics*, Vol. 101, No. 15, pp. 2469-2480.

Bogan, AA. & Thorn, KS. (1998). Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, Vol. 280, No. 1, pp. 1-9.

Böhm, HJ. (1992). The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *Journal of Computer-Aided Molecular Design*, Vol. 6, No. 1, pp. 61-78.

Brady, GP. & Stouten, PF. (2000). Fast prediction and visualization of protein binding pockets with pass. *Journal of Computer-Aided Molecular Design*, Vol. 14, No. 4, pp. 383-401.

Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; et al. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational. Chemistry*, Vol. 4, No. 2, pp. 187-217.

Brunger, AT. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, Vol. 355, No. 6359, pp. 472-475.

Brylinski, M. & Skolnick, J. (2008). Q-Dock: low-resolution flexible ligand docking with pocket-specific threading restraints. *Journal of Computational Chemistry*, Vol. 29, No. 10, pp. 1574-1588.

Carter, P. (1986). Site-directed mutagenesis. *Biochemical Journal*, Vol. 237, No. 1, pp. 1-7.

Chen, HM.; Liu, BF.; Huang, HL.; Hwang, SF. & Ho, SY (2007). SODOCK: swarm optimization for highly flexible protein–ligand docking. *Journal of Computational Chemistry*, Vol. 28, No. 2, pp. 612-623.

Chen, K.; Li, T. & Cao, T. (2006). Tribe-pso: a novel global optimization algorithm and its application in molecular docking. *Chemometrics and Intelligent Laboratory Systems*, Vol. 82, No. 1-2, pp. 248-259.

Cho, AE.; Guallar, V.; Berne, BJ. & Friesner, R. (2005). Importance of accurate charges in molecular docking: quantum mechanical/molecular mechanical (QM/MM) approach. *Journal of Computational Chemistry*, Vol. 26, No. 9, pp. 915-931.

Cho, AE. & Rinaldo, D. (2009). Extension of QM/MM docking and its applications to metalloproteins. *Journal of Computational Chemistry*, Vol. 30, No. 16, pp. 2609-2616.

Chopra, G.; Summa, CM. & Levitt, M. (2008). Solvent dramatically affects protein structure refinement. *Proceedings of the National Academy of Sciences*, Vol. 105, No. 51, pp. 20239 - 20244.

Chung, JY.; Hah, JM. & Cho, AE. (2009). Correlation between performance of QM/MM docking and simple classification of binding sites. *Journal of Chemical Information and Modeling*, Vol. 49, No. 10, pp. 2382-2387.

Connolly, ML. (1983). Analytical molecular surface calculation. *Journal of Applied Crystallography*, Vol. 16, No. 5, pp. 548-558.

Das, P.; Li, J.; Royyuru, AK. & Zhou, R. (2009). Free energy simulations reveal a double mutant avian H5N1 virus hemagglutinin with altered receptor binding specificity. *Journal of Computational Chemistry*, Vol. 30, No. 11, pp. 1654-1663.

Durbin, R. et al., (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge University Press, ISBN: 9780521629713, United Kingdom.

Evers, A. & Klabunde, T. (2005). Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha 1A Adrenergic Receptor. *Journal of Medicinal Chemistry*, Vol. 48, No. 4, pp. 1088-1097.

Ewing, TJA. & Kuntz, ID. (1997). Critical evaluation of search algorithms for automated molecular docking and database screening. *Journal of Computational Chemistry*, Vol. 18, No. 9, pp. 1175-1189.

Friesner, RA, Banks, JL, Murphy, RB, Halgren, TA, et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, Vol. 47, No. 7, pp. 1739-1749.

Giacovazzo, C. (2002). *Fundamentals of crystallography*, Oxford University Press. ISBN: 0198509588, USA.

Gräter, F.; Schwarzl, SM.; Dejaegere, A.; Fischer, S. & Smith, JC. (2005). Protein/ligand binding free energies calculated with Quantum Mechanics/Molecular Mechanics. *The Journal of Physical Chemistry B*, Vol. 109, No. 20, pp. 10474-10483.

Guex, N. & Peitsch, MC. (1997). Swiss-model and the Swiss-PDB Viewer: an environment for comparative protein modeling. *Electrophoresis*, Vol. 18, No. 15, pp. 2714-2723.

Halgren, TA.; Murphy, RB.; Friesner, RA.; Beard, HS. et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of Medicinal Chemistry*, Vol. 47, No. 7, pp. 1750-1759.

Hou, T.; Wang, J.; Li, Y. & Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, Vol. 51, No. 1, pp. 69-82.

Huang, SY. & Zou, X. (2010). Advances and challenges in protein-ligand docking. *International Journal of Molecular Sciences*, Vol. 11, No. 8, pp. 3016-3034.

Jefferys, BR.; Lawrence AK. & Sternberg, MJE. (2010). Protein folding requires crowd control in a simulated cell. *Journal of Molecular Biology*, Vol. 397, No. 5, pp. 1329-1338.

Jones, DT. (1999a) Protein secondary structure prediction based on position-specific scoring matrices *Journal of Molecular Biology*, Vol. 292, No. 2, pp. 195-202.

Jones, DT. (1999b) Genthreader: an efficient and reliable protein fold recognition method for genomic sequences *Journal of Molecular Biology*, Vol. 287, No. 4, pp. 797-815.

Jones, G.; Willett, P. & Glen, RC.; (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, Vol. 245, No. 1, pp. 43-53.

Jones, G.; Willett, P.; Glen, RC.; Leach, AR. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, Vol. 267, No. 3, pp. 727-748.

Jorgensen, WL. & Ravimohan, C. (1985). Monte carlo simulation of differences in free energies of hydration. *The Journal of Chemical Physics*, Vol. 83, No. 6, p. 3050.

Kelley, LA.; MacCallum, RM. & Sternberg, MJ. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *Journal of Molecular Biology*, Vol. 299, No. 2, pp. 499-520.

Kelley, LA. & Sternberg, MJE. (2009). Protein structure prediction on the web: a case study using the phyre server. *Nature Protocols*, Vol. 4, No. 3, pp. 363-371.

Kerzmann, A,; Fuhrmann, J.; Kohlbacher, O. & Neumann, D. (2008). BALLDock/SLICK: a new method for protein-carbohydrate docking. *Journal of Chemical Information and Modeling*, Vol. 48, No. 8, pp. 1616-1625.

Kirkwood, J. (1935). Statistical mechanics of fluid mixtures. *Journal of Chemical Physics*, Vol. 3, No. 5, pp. 300-313.

Kollman, P A,; Massova, I.; Reyes, C.; Kuhn, B. et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research*, Vol. 33, No. 12, pp. 889-897.

Kramer, B.; Rarey, M. & Lengauer, T. (1999). Evaluation of the flexx incremental construction algorithm for protein-ligand docking. *Proteins*, Vol. 37, No. 2, pp. 228-241.

Krammer, A.; Kirchhoff, PD.; Jiang, X.; Venkatachalam, CM. & Waldman, M. (2005). LigScore: a novel scoring function for predicting binding affinities. *Journal of Molecular Graphics & Modelling*, Vol. 23, No. 5, pp. 395-407.

Kuntz, ID.; Blaney, JM.; Oatley, SJ.; Langridge, R. & Ferrin, TE. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, Vol. 161, No. 2, pp. 269-288.

Laurie, ATR. (2005). Q-Sitefinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, Vol. 21, No. 9, pp. 1908-1916.

Lee, C. & Irizarry, K. (2001). The genemine system for genome/proteome annotation and collaborative data mining *IBM Systems Journal*, Vol. 40, No. 2, pp. 592-603.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology*, Vol. 226, No. 2, pp. 507-533.

Lobley, A., Sadowski, M.I. & Jones, D.T. (2009). pGenTHREADER and pDomTHREADER: New methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, 25, 1761-1767.

Lovell, SC.; Word, JM.; Richardson, JS. & Richardson, DC. (2000). The penultimate rotamer library. *Proteins: Structure, Function, and Bioinformatics*, Vol. 40, No. 3, pp. 389-408.

di Luccio, E & Koehl, P. (2011). A quality metric for homology modeling: the h-factor. *BMC Bioinformatics*, Vol. 12, p. 48.

Luthy, R.; Bowie, JU. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, Vol. 356, No. 6364, pp. 83-85.

Madej, T.; Gibrat, JF. & Bryant, SH. (1995). Threading a database of protein cores. *Proteins*, Vol. 23, No. 3, pp. 356-369.

Massova, I. & Kollman, PA. (1999). Computational alanine scanning to probe protein−protein interactions: a novel approach to evaluate binding free energies. *Journal of the American Chemical Society*, Vol. 121, No. 36, pp. 8133-8143.

McGann, MR.; Almond, HR.; Nicholls, A.; Grant, JA. & Brown, FK. (2003). Gaussian docking functions. *Biopolymers*, Vol. 68, No. 1, pp. 76-90.

McGuffin, LJ. & Jones, DT. (2003). Improvement of the genthreader method for genomic fold recognition. *Bioinformatics*, Vol. 19, No. 7, pp. 874 -881.

McMartin, C. & Bohacek, RS. (1997). QXP: powerful, rapid computer algorithms for structure-based drug design. *Journal of Computer-Aided Molecular Design*, Vol. 11, No. 4, pp. 333-344.

Miller, MD.; Kearsley, SK.; Underwood, DJ. & Sheridan, RP. (1994). FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, Vol. 8, No. 2, pp. 153-174.

Misura, KMS.; Chivian, D.; Rohl, CA.; Kim, DE. & Baker, D. (2006). Physically realistic homology models built with rosetta can be more accurate than their templates. *Proceedings of the National Academy of Sciences*, Vol. 103, No. 14, pp. 5361 -5366.

Mitchell, EP.; Sabin, C.; Šnajdrová, L.; Pokorná, M.; Perret, S.; Gautier, C.; Hofr, C.; Gilboa-Garber, N.; Koča, J.; Wimmerová, M. & Imberty, A. (2005). High affinity fucose binding of *Pseudomonas aeruginosa* lectin PA-IIL: 1.0 Å resolution crystal structure of the complex combined with thermodynamics and computational chemistry approaches. *Proteins: Structure, Function, and Bioinformatics*, Vol. 58, No. 3, pp.735-746.

Mony, L.; Krzaczkowski, L.; Leonetti, M.; Le Goff, A. et al. (2009). Structural basis of NR2B-selective antagonist recognition by n-methyl-d-aspartate receptors. *Molecular Pharmacology*, Vol. 75, No. 1, pp. 60-74.

Moreira, IS.; Fernandes, PA. & Ramos, MJ. (2007a). Computational alanine scanning mutagenesis--an improved methodological approach. *Journal of Computational Chemistry*, Vol. 28, No. 3, pp. 644-654.

Morris, GM.; Goodsell, DS.; Halliday, RS.; Huey, R.; et al. (1998). Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, Vol. 19, No. 14, pp. 1639-1662.

Morris, GM.; Huey, R.; Lindstrom, W.; Sanner, MF.; et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, Vol. 30, No. 16, pp. 2785-2791.

Najmanovich, R,; Kuttner, J,; Sobolev, V. & Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins: Structure, Function, and Genetics*, Vol. 39, No. 3, pp. 261-268.

Namasivayam, V. & Günther, R. (2007). PSO@autodock: a fast flexible molecular docking program based on swarm intelligence. *Chemical Biology & Drug Design*, Vol. 70, No. 6, pp. 475-484.

Nayeem, A.; Sitkoff, D. & Krystek Jr., S. (2006). A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. *Protein Science*, Vol. 15, No. 4, pp. 808-824.

Olson, MA.; Feig, M. & Brooks, CL. (2008). Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions. *Journal of Computational Chemistry*, Vol. 29, No. 5, pp. 820-831.

OpenEye Scientific Software, Santa Fe, New Mexico. http://www.eyesopen.com/

Pei, J.; Wang, Q.; Liu, Z.; Li, Q.; et al. (2006). PSI-Dock: towards highly efficient and accurate flexible ligand docking. *Proteins: Structure, Function, and Bioinformatics*, Vol. 62, No. 4, pp. 934-946.

Peitsch, M.C. (1995). ProMod: Automated knowledge-based protein modelling tool. *PDB Quarterly Newsletter*, 72, 4.

Prokop, M.; Adam, J.; Kříž, Z.; Wimmerová, M. & Koča, J. (2008). Triton: a graphical tool for ligand-binding protein engineering. *Bioinformatics*, Vol. 24, No. 17, pp. 1955 -1956.

Pymol. The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC. http://www.pymol.org/

Rarey, M.; Kramer, B.; Lengauer, T. & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, Vol. 261, No. 3, pp. 470-489.

Roessler, CG.; Hall, BM.; Anderson, WJ.; Ingram, WM.; et al. (2008). Transitive homology-guided structural studies lead to discovery of cro proteins with 40% sequence identity but different folds," *PNAS*, Vol. 105, No. 7, pp. 2343-2348.

Saccenti, E. & Rosato, A. (2008). The war of tools: how can NMR spectroscopists detect errors in their structures ?. *Journal of Biomolecular NMR*, Vol. 40, No. 4, pp. 251-261.

Sali, A. & Blundell, TL. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, Vol. 234, No. 3, pp. 779-815.

Sauton, N.; Lagorce, D.; Villoutreix, BO. & Miteva, MA. (2008). MS-Dock: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinformatics*, Vol. 9, No. 1, p. 184.

Schwede, T.; Kopp, J.; Guex, N. & Peitsch, MC. (2003). Swiss-model: an automated protein homology-modeling server. *Nucleic Acids Research*, Vol. 31, No. 13, pp. 3381 -3385.

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, Vol. 21, No. 7, pp. 951 -960.

Srinivasan, J.; Thomas EC.; Piotr, C.; Kollman, PA. & Case, DA. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate−DNA helices. *Journal of the American Chemical Society*, Vol. 120, No. 37, pp. 9401-9409.

Stroganov, OV.; Novikov, FN.; Stroylov, VS.; Kulkov, V. & Chilov, GG. (2008). Lead finder: an approach to improve accuracy of protein−ligand docking, binding energy estimation, and virtual screening. *Journal of Chemical Information and Modeling*, Vol. 48, No. 12, pp. 2371-2385.

Summa, CM. & Levitt, M. (2007). Near-native structure refinement using in vacuo energy minimization. *Proceedings of the National Academy of Sciences*, Vol. 104, No. 9, pp. 3177 -3182.

Taylor, RD.; Jewsbury, PJ. & Essex, JW. (2002). A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design*, Vol. 16, No. 3, pp. 151-166.

Terp, GE.; Johansen, BN.; Christensen, IT. & Jørgensen, FS. (2001). A new concept for multidimensional selection of ligand conformations (multiselect) and multidimensional scoring (multiscore) of protein-ligand binding affinities. *Journal of Medicinal Chemistry*, Vol. 44, No. 14, pp. 2333-2343.

Trott, O. & Olson, AJ. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, Vol. 31, No. 2, pp. 455-461.

Tuttle, T. (2010). Applications of QM/MM in inorganic chemistry. *Spectroscopic Properties of Inorganic and Organometallic Compounds*, pp. 87-110, Royal Society of Chemistry, Cambridge, ISBN: 9781849730853

Vásquez, M. (1996). Modeling side-chain conformation. *Current Opinion in Structural Biology*, Vol. 6, No. 2, pp. 217-221.

van Vlijmen, HWT. & Karplus, M. (1997). PDB-based protein loop prediction: parameters for selection and methods for optimization. *Journal of Molecular Biology*, Vol. 267, No. 4, pp. 975-1001.

Wang, M. & Wong, CF. (2007). Rank-ordering protein-ligand binding affinity by a Quantum Mechanics/Molecular Mechanics/Poisson-Boltzmann-Surface Area model. *The Journal of Chemical Physics*, Vol. 126, No. 2, pp. 026101.

Wang, R.; Lai, L. & Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design*, Vol. 16, No. 1, pp. 11-26.

Warren, GL.; Andrews, CW.; Capelli, AM.; Clarke, B.; et al. (2006). A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry*, Vol. 49, No. 20, pp. 5912-5931.

Warshel, A. & Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, Vol. 103, No. 2, pp. 227-249.

Weiner, PK. & Kollman, PA. (1981). AMBER: Assisted Model Building with Energy Refinement. a general program for modeling molecules and their interactions. *Journal of Computational Chemistry*, Vol. 2, No. 3, pp. 287-303.

Wells, JA. (1991). Systematic mutational analyses of protein-protein interfaces. *Methods in Enzymology*, Vol. 202, pp. 390-411.

Wuthrich, K.; (1990). Protein structure determination in solution by nmr spectroscopy. *Journal of Biological. Chemistry,* Vol. 265, No. 36, pp. 22059-22062.

Wuthrich, K. (2003). NMR studies of structure and function of biological macromolecules. *Journal of Biomolecular NMR*, Vol. 27, No. 1, pp. 13-39.

Yuan, Z.; Bailey, TL. & Teasdale, RD. (2005). Prediction of protein b-factor profiles. *Proteins: Structure, Function, and Bioinformatics*, Vol. 58, No. 4, pp. 905-912.

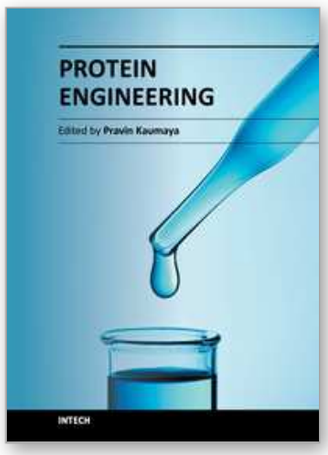
Yuriev, E.; Agostino, M. & Ramsland, PA. (2011). Challenges and advances in computational docking: 2009 in review. *Journal of Molecular Recognition: JMR*, Vol. 24, No. 2, pp. 149-164.

Zhang, H. (2002). Protein Tertiary Structures: Prediction from Amino Acid Sequences, *Encyclopedia of Life Sciences,* Macmillan Publishers Ltd, Nature Publishing Group, England.

Zhu, J.; Fan, H.; Periole, X.; Honig, B. & Mark, AE. (2008). Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins: Structure, Function, and Bioinformatics*, Vol. 72, No. 4, pp. 1171-1188.

Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, BS. & Johnson, AP. (2006). eHITS: an innovative approach to the docking and scoring function problems. *Current Protein & Peptide Science*, Vol. 7, No. 5, pp. 421-435.

Zwanzig, R. (1954). High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics,* Vol. 22, No. 8, pp. 1420-1426.



Protein Engineering

Edited by Prof. Pravin Kaumaya

ISBN 978-953-51-0037-9

Hard cover, 344 pages

Publisher InTech

Published online 24, February, 2012

Published in print edition February, 2012

A broad range of topics are covered by providing a solid foundation in protein engineering and supplies readers with knowledge essential to the design and production of proteins. This volume presents in-depth discussions of various methods for protein engineering featuring contributions from leading experts from different counties. A broad series of articles covering significant aspects of methods and applications in the design of novel proteins with different functions are presented. These include the use of non-natural amino acids, bioinformatics, molecular evolution, protein folding and structure-functional insight to develop useful proteins with enhanced properties.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Sushil Kumar Mishra, Gabriel Demo, Jaroslav Koča and Michaela Wimmerová (2012). In Silico Engineering of Proteins That Recognize Small Molecules, Protein Engineering, Prof. Pravin Kaumaya (Ed.), ISBN: 978-953-51-0037-9, InTech, Available from: http://www.intechopen.com/books/protein-engineering/in-silico-engineering-of-proteins-that-recognize-small-molecules

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen