Chapter-3
3 PROTEIN STRUCTURE PREDICTIONS

3.1 Introduction to protein structure

Proteins are an important class of biological macromolecules present in all biological organisms that are made up of such elements as carbon, hydrogen, nitrogen, oxygen, and sulphur. All proteins are polymers of amino acids composed by a sequence of 20 different L-α-amino acids that is also referred as residues. The best sources of protein are beef, poultry, fish, eggs, dairy products, nuts, seeds and legumes like black beans and lentils. The name protein derives from the Greek "Protos", meaning "first" or "foremost". All proteins, whether from the most ancient lines of bacteria or from the most complex forms of life, are constructed from the same ubiquitous set of 20 amino acids, covalently linked in characteristic linear sequences. From these building blocks different organisms can make such widely diverse products as enzymes, hormones, antibodies, transporters, muscle, the lens protein of the eye, feathers, spider webs, rhinoceros horn, milk proteins, antibiotics, mushroom poisons e.t.c having distinct biological activities. Aminoacids are classified on the basis of the polarity and charge (at pH 7) of their R groups. The nonpolar, aliphatic class includes alanine, glycine, isoleucine, leucine, methionine and valine. Phenylalanine, tryptophan, and tyrosine have aromatic side chains and are relatively hydrophobic. The polar uncharged class includes asparagines, cysteine, glutamine, proline, serine, and threonine. The negatively charged (acidic) amino acids are aspartate and glutamate; the positively charged (basic) ones are arginine, histidine, and lysine. 

Proteins are integral to the formation of DNA, a molecule that contains genetic codes for inheritance and of hormones. To be able to perform their biological function, proteins fold into one, or more specific spatial conformations which are driven by a number of noncovalent interactions such as hydrogen bonding, ionic interactions, vanderwaal forces and hydrophobic packing. In order to understand the functions of proteins at a molecular level, it is often necessary to determine the three dimensional structures of proteins.

3.2 Protein structures are generally described at four levels

3.2.1 Primary structure (1°)

The unique sequence of the different amino acids is called the primary structure of the peptide or protein. Amino acids in a protein can be joined covalently through peptide bonds formed between amino (-NH) group of one amino acid to another.
amino acid of carboxylic (-Coo) group to form peptides and proteins. The primary structure of a protein is determined by the gene corresponding to the protein. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as Edman degradation or tandem mass spectrometry. Often, it is read directly from the sequence of the gene using the genetic code. "Misspellings" in the order of amino acids during their transcription and translation causes change in the nature of protein that is drastic enough to life threatening if the protein has critical function. For example, substitution of single glutamic acid residue of position sixth of the Beta chains in hemoglobin by a valine residue causes "sickle cell anemia" is a disease characterized by the "Sickling of red cells" or red cells induced by lowered oxygen tension.

3.2.2 Secondary structure (2°)

Secondary structure is the regular arrangement of amino acid residues in a segment of a polypeptide chain, in which each residue is spatially related to its neighbors in the same way. The most common secondary structures are the α helix, and the β conformation. Using fundamental chemical principles and a few experimental observations, Pauling and Corey predicted the existence of these secondary structures in 1951, several years before the first complete protein structure was elucidated. The simplest arrangement of the polypeptide chain could assume with its rigid peptide bonds (but other single bonds free to rotate) is a helical structure called the α helix (Pauling and Corey 1951) (Fig 3.1a.) In this structure the polypeptide backbone is tightly wound around an imaginary axis drawn longitudinally through the middle of the helix, and the R groups of the amino acid residues protrude outward from the helical backbone. The repeating unit is a single turn of the helix, which extends about 5.4 Å along the long axis. The amino acid residues in an α helix have conformations with \( \psi = -45^\circ \) to \(-50^\circ \) and \( \phi = -60^\circ \), and each helical turn includes 3.6 amino acid residues. The helical twist of the \( \alpha \) helix found in all proteins is right-handed. the \( \alpha \) helix proved to be the predominant structure in α keratins. More generally, about one-fourth of all amino acid residues in polypeptides are found in \( \alpha \) helices, the exact fraction varying greatly from one protein to the next. The turns of \( \alpha \) helix are stabilized by hydrogen bonding between every fourth amino acid in the chain. Further model-building experiments have shown that \( \alpha \) helix can form in polypeptides consisting
of either L or D-amino acids. However, all residues must be of one stereo isomeric series; a D-amino acid will disrupt a regular structure consisting of L-amino acids, and vice versa. Naturally occurring L-amino acids can form either right or left-handed α helices, but extended left-handed helices have not been observed in proteins. A second type of repetitive structure, called the β conformation which is a more extended conformation of polypeptide chains, and its structure has been confirmed by x-ray analysis. In the β conformation, the backbone of the polypeptide chain is extended into a zigzag rather than helical structure (Fig.3.1b).

Fig 3.1 Diagrammatic representation of a) α helix and b) β sheet
The zigzag polypeptide chains can be arranged side by side to form a structure resembling a series of pleats, called a β sheet. Hydrogen bonds in a β sheet are formed between adjacent segments of polypeptide chain. The individual segments that form a β sheet are usually nearby on the polypeptide chain, but can also be quite distant from each other in the linear sequence of the polypeptide; they may even be segments in different polypeptide chains. The R groups of adjacent amino acids protrude from the zigzag structure in opposite directions, creating the alternating pattern seen in the side views. The adjacent polypeptide chains in a β sheet can be either parallel or anti parallel (having the same or opposite amino-to-carboxyl orientations, respectively). The structures are somewhat similar, although the repeat period is shorter for the parallel conformation (6.5 Å, versus 7 Å for anti parallel) and the hydrogen bonding patterns are different. Amino acid residues in the β-conformation have negative Φ angles and the Ψ angles are positive. Typical values are Φ = -140 degrees and Ψ =
130 degrees. In contrast, \( \alpha \)-helical residues have both \( \Phi \) and \( \Psi \) negative.

**Ramachandran plot**

Ramachandran plot is one of the simplest and most sensitive means for assessing the quality of a protein model in the absence of experimental data (Ramachandran et al., 1968). With the advent of the program Procheck, Ramachandran plots have gained somewhat in popularity. Procheck divides the Ramachandran plot into four types of area: most favoured, additional allowed, generously allowed and disallowed. A typical good model should not only have few residues within the disallowed regions, but also many in the most favoured regions. The plot shows the possible conformations of phi and psi angles for a polypeptide. Every type of secondary structure can be completely described by the bond angles \( \Phi \) and \( \psi \) at each residue. As shown by a Ramachandran plot, the \( \alpha \) helix and \( \beta \) conformation fall within a relatively restricted range of sterically allowed structures (Fig. 3.2). Most values of \( \Phi \) and \( \psi \) taken from known protein structures fall into the expected regions, with high concentrations near the \( \alpha \) helix and \( \beta \) conformation values around (-62,-41) and (-120, +120). The only amino acid residue often found in a conformation outside these regions is glycine. Due to lack of C\( \beta \) atom, it is more flexible and can therefore cover a wider area of the Ramachandran map. Proline having its side chain connected to N atom has fewer allowed torsion angle. The remaining 18 amino acids are usually placed in the same map as the differences in backbone conformation are relatively small.
An Uparidiei CoJa^en triple-

In the Fig.3.2 white areas correspond to conformations where atoms in the polypeptide come closer than the sum of their vanderwaal's radius. These regions are sterically disallowed for all amino acids except glycine which is unique in that it lacks a side chain. The dark blue regions correspond to conformations where there are no steric clashes, i.e. these are the allowed regions namely the alpha-helical and beta-sheet conformations. The light blue areas show the allowed regions if slightly shorter vanderwaal's radi are used in the calculation, i.e the atoms are allowed to come a little closer together. This brings out an additional region which corresponds to the left-handed alpha helix.

3.2.3 Tertiary structure (3°)

Tertiary structure refers to the complete three-dimensional structure of the polypeptide units of a given protein. Secondary structures of proteins often constitute distinct domains. Therefore, tertiary structure also describes the relationship of different domains to one another within a protein. The interactions of different domains are governed by several forces; these include hydrogen bonding, hydrophobic interactions, electrostatic interactions and vanderwaal's forces. Tertiary structure is often stabilized by disulfide bonds between adjacent
cysteine in different regions of the protein. For example, the tertiary structure of Ribonuclease contains four disulfide bonds located at specific sites. The stability of the tertiary structure of proteins is destroyed by toxic heavy metals such as Mercury. Concentrations of mercury in the environment may result in the displacement of hydrogen on the sulphur atoms (SH), thereby blocking functional disulfide bonds. Several other weak, noncovalent interactions also help to stabilize the tertiary structure. These noncovalent interactions can be disrupted by heating a protein or exposing it to extreme pH (acidity or alkalinity), which alters the charge of polar groups on the amino acids. Such disruptions cause the protein to unfold, often exposing hydrophobic groups and leading to precipitation (clumping together) of the protein. If these disruptive factors are removed, some proteins can refold to their original conformation. This ability to refold confirms that protein folding is a self assembly process that is dependent upon the sequence of amino acids.

3.2.4 Quaternary structure (4°)
The quaternary structure is the interaction between several chains of peptide bonds. The individual chains are called subunits. The individual subunits are usually not only covalently connected, but might be connected by a disulfide bond. Not all proteins have quaternary structure, since they might be functional as monomers. Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers. Specifically it would be called a dimer if it contains two subunits, a trimer if it contains three subunits and a tetramer if it contains four subunits. The subunits are usually related to one another by symmetry axes, such as a fold axis in a dimer. Multimers made up of identical subunits may be referred to with a prefix of "homo-" (e.g homo tetramer) and those made up of different subunits may be referred to with a prefix of "hetero-" (e.g. two alpha and two beta chains of hemoglobin). All the four levels of protein structure are shown in the Fig 3.3.
3.3 Protein structure prediction by computational methods

Understanding the mechanism of protein function generally requires knowledge of protein three-dimensional structure, which is ultimately determined by protein sequence. Protein structure determination using experimental methods such as X-ray crystallography or NMR spectroscopy is time consuming and not successful with all proteins, especially membrane proteins. Currently there are about two million protein sequences in Swissprot and TrEMBL (http://us.expasy.org/sprot/), among them about 30,000 proteins have had their structures solved experimentally (http://www.rcsb.org/pdb/). Although the rate of experimental structure determination will continue to increase, the number of newly discovered sequences grows much faster than the number of structures solved. The huge gap between the number of available sequences and experimentally solved protein structures could possibly be resolved by computational methods.

Computational protein structure prediction can be divided into two extreme camps:

- ab-initio method
- Homology modeling.

3.3.1 Ab initio prediction

Ab initio prediction (Bonneau and Baker, 2001) was based on amino acid sequence of the protein with unknown structure, which is the most difficult method. The aim is to predict the structure based on physical principles and general studies. The ab initio prediction methods consist of modeling all the
energetics involved in the process of folding, and then in finding the structure with lowest free energy. This approach is based on the 'thermodynamic hypothesis', which states that the native structure of a protein is the one, for which the free energy achieves the global minimum. Standard search techniques such as genetic algorithms, Montle Carlo and stimulated annealing are usually applied to explore the conformational space of proteins. The resolution of current ab initio structure prediction techniques may be sufficient for genome annotation; however, it is clearly not yet precise enough for detailed studies such as docking and drug design (Hardin et al., 2002).

3.3.2 Brief History of homology modeling studies

The first modeling studies, carried out in the late 1960 and early 1970s, were based upon the construction of wire or plastic models. The later studies were performed using interactive computer graphics. Browne et al., (Browne et al., 1969) published the first report on homology modeling. They modeled bovine α-lactalbumin on the known 3D structure of hen egg white lysozyme. Later on Warme et al., (Warme et al., 1974) produced a model for α-lactalbumin on the basis of the crystal structure of lysozyme. These models were constructed by taking the existing coordinates of the known structure, and mutating side chains not identical in the protein to be modeled. This approach to protein modeling is still employed today with considerable success, especially when the proteins are similar (May and Blundell, 1994). McLachlan and Shotton (McLachlan and Shotton, 1971) modeled α-lytic proteinase of the fungus Myxobacter 495 on the basis of the structure of the mammalian chymotrypsin and elastase. The modeling was a difficult task because the sequence identity between the one to be modeled and the known structures was of the order of 18%. Subsequently, Brayer et al., (Brayer et al., 1979) determined the crystal structure of α-lytic proteinase and compared the X-ray structure with the homology model. They found that although segment of both domains of the model were built correctly, misalignment of the sequences led to local errors. Among the aspartic proteinase the first models were constructed for renin and renin inhibitor complexes using the 3D structure of the distantly related fungal proteinases. Later on, the homology models for renin were built using the structures of mammalian aspartic proteases, pepsin and chymosin (Hutchins and Greer, 1991). Comparison of the renin
models constructed from fungal and mammalian enzymes revealed that errors in the models arose from the differences in the arrangement of helices and strands between the mammalian and fungal aspartic proteinases, as well as the rather different variable regions. Nevertheless, the active site of the renin was modeled reasonably correctly. In the early eighties, manual homology modeling was facilitated by manipulation of protein molecules on the graphics terminal that was made possible by computer programs such as FRODO (Jones, 1978). Since the mid-1980s, a large number of homology models of proteins with different folds and functions have been reported in the literature.

3.3.3 Comparative modeling

It also includes threading (Domingues et al., 2000), it compares to a target sequence against a library of structural templates, producing a list of scores. The scores are then ranked and the fold with the best score is assumed to be the one adopted by the sequence. Homology or comparative modeling (Blundell et al., 1987, Martí-Renom et al., 2000) is a technique to create a 3D structure of a target protein in atomic detail on the basis of a known template structure from a homologous protein that was previously determined with the help of NMR or X-ray experiments. This is possible if both sequences are very similar because analysis of experimental structures indicate that one or more proteins sharing an amino acid identity of more than 30% to the target protein with the unknown structure can be used as templates to model a new structure (Martí-Renom et al., 2000).

More applications of homology modeling are feasible with increasing accuracy of the model, which in turn depends significantly on the sequence identity between the target and the template structure. Homology modeling has assumed an increasingly important role in protein structure prediction. This is because many protein sequences are evolutionarily related, and thus can be classified into different families. Proteins in the same families frequently have noticeable similarities and thus share three-dimensional architecture, which allows a structural description of all proteins in a family even when only the structure of a single member is known. This evolutionary relationship provides the rationale for structural genomics, a systematic and large-scale effort towards structural characterization of all proteins, where a representative protein in each family is chosen to be solved experimentally with the rest reliably predicted by a homology modeling method (Goldsmith Fischman et al., 2003). Even though ab initio method
can not solve protein folding problem in the foreseeable future, structure prediction will nevertheless be solved by homology modeling method anyway with the completion of structural genomics project, which looks like an attainable goal in the next 10 years. In fact, if we assume that protein structure is of global energy minimum, homology modeling is a simple scheme to search the conformation space by minimally disturbing those existing solutions, i.e., the experimentally solved structures. The obvious advantage is that the homology modeling technique relaxes the stringent requirement of force field and enormous conformation searching, because it dispenses with the calculation of a physical chemistry force field and replaces it, in large part, with the counting of sequence identities.

Given a protein sequence, homology modeling usually consists of the following four steps (Fiser and Sali 2003, Sanchez 1997) 1) identify the homologue of known structure from the Protein data bank 2) align the query sequence to the template structure 3) build the model based on the alignment; 4) assess and refine the model. When the sequence identity is above 40%, the alignment is straightforward, there are not many gaps, and 90% of main-chain atoms could be modeled with an RMSD (root-mean-square distance) error of about 1 Å (20). In this range of sequence identity, the structural difference between proteins mainly arises from loops and side-chains. Proteins with such a sequence similarity are supposed to be evolved divergently from a common ancestor and are therefore typically classified as homologous. When the sequence identity is about 30–40%, obtaining correct alignment becomes difficult, where insertions and deletions are frequent. For sequence similarity in this range, 80% of main-chain backbone atoms can be predicted to RMSD of 3.5Å, while the rest of residues are modeled with larger errors, especially in the insertion and deletion regions (Harrison et al., 1995). Even in correctly aligned regions, loop modeling and side-chain placement pose difficulties (Bower 1997). When the sequence similarity is below 30%, the main problem becomes the identification of the homologue structures and alignment becomes much more difficult. Approximately 57% of all known sequences have at least one domain that is related to at least one protein of known structure (Pieper et al., 2002). The probability of finding a related known structure for a randomly selected sequence from a genome ranges from 30% to 65%, since a few genomes have received more research attention than others.
The percentage is steadily increasing because more distinct folds are discovered each year, and because the number of different structural folds that proteins adopt is limited. Current estimates suggest that there are between 1000 and 5000 folds in the universe of compact globular proteins, with about 200 new folds realized annually from the structure deposition (Brenner et al., 1997). Currently, over 1.1 million proteins can readily have at least one of their domains reliably predicted with homology modeling methods. Given the rate of experimental structural determination at approximately 6000 proteins each year (http://www.rscb.org), it is arguable that homology modeling has already saved up to hundreds of years of human effort. In the next 10 years, structural genomics will possibly discover all protein distinct folds in Nature, making homology modeling applicable to almost any protein sequence. The usefulness of homology modeling is ever increasing when more proteins can be predicted with higher accuracy.

Homology modeling is a multi-step process that can be summarized as below

- Template selection and Sequence alignment
- Model building by assembly of rigid bodies
- Modeling by segment matching
- Modeling by Modeller
- Loop modeling
- Side chain modeling
- Model Refinement
- Model Assessment

3.3.3.1 Template selection and sequence alignment

At the beginning of modeling it is necessary to identify all appropriate templates that are sufficiently related to the target protein. The main features used to detect protein-protein relationships are sequence and structure. Hence, structures and corresponding information data are available at databases like the Protein structure Data Bank (PDB) (Berman et al., 2000), SCOP (Murzin et al., 1995) and CATH (Orengo et al., 1997; Pearl et al., 2000) to enable searching for suitable templates for assigning a 3D fold to the target sequence. The number of different methods that have to be used to choose the templates relies on the number of available homologs in the PDB and their overall sequence identity to the target sequence. If many structures with highly homologous sequences can be found, pair-wise alignments between the target and each template can be done using tools such as BLAST (Altschul et al., 1990) or FASTA (Pearson 1998) against the PDB. At lower sequence similarity techniques like multiple sequence alignments
(Fig 3.4) become more and more important. Here, programs like PSI-BLAST are used to expand the set of homologs to the target protein. Multiple alignments elucidate for instance relationships between several sequences even if their identity is very low. They are especially indispensable if only a few structures are available having a remote relationship to the target sequence or to make a special sequence pattern like sequence repeats visible.

Fig. 3.4 Example of a multiple alignment of some proteins from the SH2 family made with ClustalX. The stars at the top correspond to conserved residues among all present proteins and the dots indicate conservative substitutions between amino acids with similar properties like hydrophobicity (also visible in colour code). The bars at the bottom show a value for the similarity of the amino acids located at the same position (identity=100 %). The gaps within the 1JWO sequence correspond to deletions of amino acids located in loop regions if compared to the other homologs.

If finally the sequence identity drops beneath 30% one can also apply threading tools for fold recognition (Xu and Xu 2000; Shi et al., 2001). Instead of optimising a sequence alignment, here, comparisons are done with all possible templates and a final fold is assigned after comparing many rough models. The threading methods consider not only sequence but also empirical geometric structure patterns into account. The target sequence is threatened over a representative set of proteins from the PDB and with a scoring function the compatibility of the sequence to the structure is estimated. The scoring function itself depends on probability distributions calculated from observed residue-residue distances among all known structures and distances within their sequences. From these distributions a total system threading energy value is calculated to evaluate the
best structure predictions with lowest energies. Threading is also very helpful if
certain parts of the target sequence cannot be covered by any sequence within
the multiple alignments, so that the information about the putative fold must come
from other sources.

3.3.3.2 Alignment optimization

After template selection programs like ClustalX (Thompson et al., 1997) or Swiss-
pdb viewer (Boeckmann et al., 2003) can be applied to make an optimised final
sequence alignment that can then be translated into a structure. ClustalX is a
program to carry out and analyze multiple alignments. Swiss-pdb viewer is a
powerful graphical interface to apply homology modelling. Among others Swiss-
pdb viewer offers features for structural superposition's corresponding to
sequence alignments, structural manipulations in terms of mutations and energy
minimisations as well as for loop and side-chain modelling. While in ClustalX only
the pure sequences can be treated, the Swiss-pdb viewer offers the possibility to
adapt the alignment "by hand" in accordance to the meanwhile displayed structure
in sequence regions of low homology were the correct alignment is uncertain.
Alignment optimization means therefore to prepare locally another more sensitive
multiple alignments and a Phylogenetic tree in addition using ClustalX. This
increases the accuracy compared to using the more rough tools offered in the
World Wide Web, which are more useful for pre-selecting a broader ensemble of
templates. ClustalX divides the ensemble in sequence classes or suggested
protein subfamilies from where one representative template, maybe with closest
sequence to the target, can be chosen. This is an important step to iteratively
delete redundant sequences to optimise a multiple alignment with the reduced
number of sequences as in a numerical calculation the possible error can
increase with the number of redundant variables (sequences). In ClustalX,
alignments are performed using a substitution matrix (Fig 3.5) where every
component stands for the tenfold logarithm of an empirically determined
probability for exchanging one amino acid by another. A mutation from isoleucine
to valine is for instance very likely if one counts such mutations within homologous
proteins and also their abilities to build and conserve β strands are very similar.
Therefore a high positive value of 4 is allocated to this mutational event. In
contrast, a single (point) mutation can lead to completely abolishing of function, in
particular if for instance a huge alkaline or acidic amino acid (arginine or aspartic
acid) was substituted by a small amino acid (glycine). For this reason such substitutions are as a consequence assigned negative values in the matrix.

```
C 12
S 0 2
T -7 1 3
P -3 1 0 4
A -2 1 1 2
G -3 1 0 -1 1 5
N -4 1 0 -1 0 0 2
D -5 0 0 -2 0 1 2 4
E -5 0 0 -1 0 0 1 2 6
Q -5 1 -1 0 0 -1 1 2 4 4
H -3 -1 -1 0 -1 -2 2 1 1 3 6
R -4 0 -1 0 -2 -3 0 -1 -1 1 2 6
K -5 0 0 -1 -1 -2 1 0 0 1 0 3 5
M -5 -2 -1 -2 -1 -3 -2 -3 -2 -1 -2 0 0 6
I -2 -1 0 -2 -1 -3 -2 -2 -2 -1 -2 -2 2 5
L -6 -3 -2 -3 -2 -4 -3 -4 -3 -2 -2 -1 -3 4 2 6
V -2 -1 0 -1 0 -1 -2 -2 -2 -2 2 2 4 2 4
F 4 5 3 -5 -4 -5 -4 -6 -5 -5 -4 -4 -5 0 1 2 -1 9
Y 0 -3 -3 -5 -3 -5 -2 -4 -4 -4 0 -4 -4 -2 -1 -1 -2 7 10
W -8 -4 -4 -6 -7 -4 -4 -7 -7 -5 -3 -2 -3 -4 -3 -2 -6 0 0 17
```

Fig. 3.5: The PAM250 similarity matrix (Altschul 1991) assigns a certain exchange value to every possible amino acid substitution. They are proportional to the tenfold logarithm of the observed mutation rate among sequences with an overall identity of more than 20% divided by the mutation rate expected from random amino acid frequencies. High positive and negative matrix values correspond therefore to high and low possibilities for the respective substitution within homologous proteins.

Performing an alignment means now to map the template onto the target sequence such that every amino acid from the target is allocated to one from the template in a way that the sum of all exchange values derived from the substitution matrix becomes maximal (Fig 3.5). Due to the fact that during evolution of proteins deletions or insertions of one or more amino acids are very common gaps between two amino acids are also allowed in the alignment where amino acids from one protein have no assigned partners in the other.

3.3.3.3 Model building

3.3.3.3.1 Modeling by assembly of rigid bodies

The first and still widely used approach in comparative modeling is to assemble a model from a small number of rigid bodies obtained from the aligned protein structures (Browne et al., 1969; Greer, 1981; Blundell et al., 1987). The approach is based on the natural dissection of the protein structures into conserved core regions, variable loops that connect them and side chains that decorate the backbone. For example, the following semi automated procedure is implemented...
in the computer program COMPOSER (Sutcliffe et al., 1987a). First, the template structures are selected and superposed. Second, the "framework" is calculated by averaging the coordinates of the C\textsuperscript{\alpha} atoms of structurally conserved regions in the template structures. Third, the main-chain atoms of each core region in the target model are obtained by superposing the core segment, from the template whose sequence is closest to the target, on the framework. Fourth, the loops are generated by scanning a database of all known protein structures to identify the structurally variable regions that fit the anchor core regions and have a compatible sequence (Topham et al., 1993). Fifth, the side chains are modeled based on their intrinsic conformational preferences and on the conformation of the equivalent side chains in the template structures (Sutcliffe et al., 1987b). Finally, the stereochemistry of the model is improved either by a restrained energy minimization or a molecular dynamics refinement. The accuracy of a model can be somewhat increased when more than one template structure is used to construct the framework and when the templates are averaged into the framework using weights corresponding to their sequence similarities to the target sequence (Srinivasan and Blundell, 1993). Possible future improvements of modeling by rigid-body assembly include incorporation of rigid body shifts, such as the relative shifts in the packing of helices and \beta-sheets (Nagarajaram et al., 1999). Two other programs that implement this method are 3D-JIGSAW (Bates et al., 2001) and SWISSMODEL (Schwede et al., 2003).

### 3.3.3.3.2 Modeling by segment matching or coordinate reconstruction

The basis of modeling by coordinate reconstruction is the finding that most hexapeptide segments of protein structure can be clustered into only 100 structurally different classes (Claessens et al., 1989). Thus, comparative models can be constructed by using a subset of atomic positions from template structures as guiding positions to identify and assemble short, all-atom segments that fit these guiding positions. The guiding positions usually correspond to the C\textsuperscript{\alpha} atoms of the segments that are conserved in the alignment between the template structure and the target sequence. The all-atom segments that fit the guiding positions can be obtained either by scanning all known protein structures, including those that are not related to the sequence being modeled (Claessens et al., 1989; Holm and Sander, 1991), or by a conformational search restrained by energy function. This method can construct both main-chain and side-chain atoms, and can also model
unaligned regions (gaps). It is implemented in the program SegMod. Even some side-chain modeling methods and the class of loop construction methods based on finding suitable fragments in the database of known structures (Jones and Thrup, 1986) can be seen as segment-matching or coordinate-reconstruction methods.

3.3.3.3 Modeling by satisfaction of spatial restraints

The methods in this class begin by generating many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. The procedure is conceptually similar to that used in determination of protein structures from NMR-derived restraints. The restraints are generally obtained by assuming that the corresponding distances between aligned residues in the template and the target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom-atom contacts that are obtained from a molecular mechanics force field. The model is then derived by minimizing the violations of all the restraints. This optimization can be achieved either by distance geometry or real-space optimization. For example, an elegant distance geometry approach constructs all-atom models from lower and upper bounds on distances and dihedral angles.

3.3.3.4 Comparative protein structure modeling by MODELLER

MODELLER, the authors' own program for comparative modeling, belongs to this group of methods (Sali and Overington, 1994; Fiser et al., 2000). MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints. The program was designed to use as many different types of information about the target sequence as possible. In the first step of model building, distance and dihedral angle restraints on the target sequence are derived from its alignment with template 3-D structures. The form of these restraints was obtained from a statistical analysis of the relationships between similar protein structures. The analysis relied on a database of 105 family alignments that included 416 proteins of known 3D structure (Sali and Overington, 1994). By scanning the database of alignments, tables quantifying various correlations were obtained, such as the correlations between two equivalent C$^\alpha$-C$^\beta$ distances, or between equivalent main-chain dihedral angles from two related proteins (Sali and Blundell, 1993). These relationships are expressed as
conditional probability density functions (pdfs), and can be used directly as spatial restraints. For example, probabilities for different values of the main-chain dihedral angles are calculated from the type of residue considered, from main-chain conformation of an equivalent residue, and from sequence similarity between the two proteins. Another example is the pdf for a certain Ca-Ca distance given equivalent distances in two related protein structures. An important feature of the method is that the form of spatial restraints was obtained empirically, from a database of protein structure alignments.

In the second step, the spatial restraints and the CHARMM22 force-field terms enforcing proper stereochemistry are combined into an objective function. The general form of the objective function is similar to that in molecular dynamics programs, such as CHARMM22. The objective function depends on the Cartesian coordinates of 10,000 atoms (3D points) that form the modeled molecules. For a 10,000-atom system, there can be on the order of 200,000 restraints. The functional form of each term is simple; it includes a quadratic function, harmonic lower and upper bounds, cosine, a weighted sum of a few Gaussian functions, Coulomb law, Lennard-Jones potential, and cubic splines. The geometric features presently include a distance, an angle, a dihedral angle, a pair of dihedral angles between two, three, four, and eight atoms, respectively, the shortest distance in the set of distances, solvent accessibility, and atom density that is expressed as the number of atoms around the central atom. Some restraints can be used to restrain pseudo-atoms, e.g. the gravity center of several atoms. Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method (Braun and Go, 1985), employing methods of conjugate gradients and molecular dynamics with simulated annealing (Clore et al., 1986). Several slightly different models can be calculated by varying the initial structure, and the variability among these models can be used to estimate the lower bound on the errors in the corresponding regions of the fold.

3.3.3.4.1 Restraints derived from experimental data
Because the modeling by satisfaction of spatial restraints can use many different types of information about the target sequence, it is perhaps the most promising of all comparative modeling techniques. One of the strengths of modeling by satisfaction of spatial restraints is that restraints derived from a number of different
sources can easily be added to the homology-derived restraints. For example, restraints could be provided by rules for secondary-structure packing, analyses of hydrophobicity and correlated mutations (Taylor et al., 1994), empirical potentials of mean force (Sippl, 1990), nuclear magnetic resonance (NMR) experiments, cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis, and intuition, among other sources. Especially in difficult cases, a comparative model could be improved by making it consistent with available experimental data and/or with more general knowledge about protein structure.

### 3.3.3.4.2 Relative accuracy, flexibility, and automation.

Accuracies of the various model-building methods are relatively similar when used optimally. Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy, especially for models based on low sequence identity to the templates. However, it is important that a modeling method allow a degree of flexibility and automation to obtain better models more easily and rapidly.

Table 3.1 shows the most widely used model building programs such as MODELLER, SEGMOD/ENCAD, SWISS-MODEL, 3D-JIGSAW, NEST and BUILDER. The three modeling programs, MODELLER, NEST and SEGMOD/ENCAD, perform better than the others.

<table>
<thead>
<tr>
<th>Programs</th>
<th>Availability</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEST</td>
<td><a href="http://trantor.bioc.columbia.edu/programs/ackal">http://trantor.bioc.columbia.edu/programs/ackal</a></td>
<td>Artificial evolution</td>
</tr>
<tr>
<td>COMPOSER</td>
<td><a href="http://www.cryst.bioc.cam.ac.uk/">http://www.cryst.bioc.cam.ac.uk/</a></td>
<td>Rigid —body assembly</td>
</tr>
<tr>
<td>TRIPOS</td>
<td><a href="http://www.tripos.com">http://www.tripos.com</a></td>
<td>Rigid —body assembly</td>
</tr>
<tr>
<td>CONGEN</td>
<td><a href="http://www.congenomics.com/congen/congen.toc.html">http://www.congenomics.com/congen/congen.toc.html</a></td>
<td>Rigid —body assembly</td>
</tr>
<tr>
<td>MODELLER</td>
<td><a href="http://guitar.rockefeller.edu/modeller/modeller.html">http://guitar.rockefeller.edu/modeller/modeller.html</a></td>
<td>Spatial restraints</td>
</tr>
</tbody>
</table>

Table 3.1 Homology Modeling Programs
Detailed analysis of these homology modeling programs revealed some interesting differences. For example, using a 1.4 GHz AMD XP processor, NEST needs 17s on average to build a model, while SEGMOD needs 6s and MODELLER takes 43 to 430s in MODELLER 6VZ and MODELLER6V2-10 respectively.

### 3.3.5 Loop modeling

Loop modeling is an especially important aspect of comparative modeling in the range from 30% to 50% sequence identity. In this range of overall similarity, loops among the homologus vary while the core regions are still relatively conserved and aligned accurately. Loop modeling can be seen as a mini protein folding problem, because the correct conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, loops are generally too short to provide sufficient information about their
local fold. Even identical deca-peptides in different proteins do not always have the same conformation (Kabsch and Sander, 1984; Mezei, 1998). Some additional restraints are provided by the core anchor regions that span the loop and by the structure of the rest of the protein that cradles the loop. Although many loop-modeling methods have been described, it is still challenging to correctly and confidently model loops longer than "8 to 10 residues" (Fiser et al., 2000, Jacobson et al., 2004). Gaps usually found during alignment, correspond to loop segments which indicates an extension of either target or template sequence. Especially in gap regions the structural information is disturbed and as a consequence loops have then to be built separately from the main-chain superposition. Loop modeling is of high importance since in evolution most mutations occur here and contribute often to active binding or recognition sites determining functional specificity of the protein. Insertions of amino acids provide loop extension and deletions lead to abbreviations of the loop segment and together with point mutations all these alterations generate a large number of different loops among all proteins. Even highly homologous proteins with very similar sequences can have different loops, each with a unique main-chain conformation. Thus, sequence similarity can hardly be used and other techniques are needed to predict the main chain conformations of loops. In Swiss-pdb viewer a loop structure library offering several hundreds of different loops enables a simple prediction procedure although validation of accuracy is rather complicated. In contrast, loop prediction in MODELLER does not rely on a database of known structures. In fact, the ab initio loop prediction method in MODELLER optimizes the positions of all non-hydrogen atoms of a loop in a fixed environment (Fiser and Sali 2003). The optimization relies on a protocol consisting of the conjugate gradient minimisation and molecular dynamics simulation with simulated annealing. The pseudo-energy function contains terms from a molecular mechanics force field CHARMM-22 as well as restraints for bonds, angles, some dihedral angles and improper dihedral angles based on statistical distributions derived from known protein structures (MacKerell et al., 1998). During simulation the main-chain and side chain dihedral angles as well as non-bonded atom pairs are restrained by statistical potentials extracted from many known protein structures (Sali and Blundell 1993; Sali and Overington 1994; Melo and Feytmans 1997; Fiser et al., 2000). The modeling protocol generates a number of
independently optimized conformations, starting with random initial conformations. The final loop prediction is the optimized conformations that have the lowest pseudo energy. Table 3.2 shows some loop modeling software that can be easily obtained from the web. Compared with database scanning method, most ab-initio loop prediction programs are very slow. For example, the loop prediction program by Fiser et al., 2000 requires about 40 hours for an 8-residue loop; while LOOPY needs only about 10 minutes with good accuracy (Xiang 2001) recent participation in CASP6 (Critical Assessment of Structure Prediction), 2004, LOOPY was ranked the third in the category of loop assessment.

Table 3.2: Loop Modeling Programs

<table>
<thead>
<tr>
<th>Programs</th>
<th>Availability</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOOPY</td>
<td><a href="http://trantor.bioc.columbia.edu/programs.html">http://trantor.bioc.columbia.edu/programs.html</a></td>
<td>Colony energy with ab-initio conformation sampling and torsional space minimization</td>
</tr>
<tr>
<td>PLOP</td>
<td><a href="http://francisco.compbio.ucsf.edu/-jacobson/plop_manual/plop_overview.html">http://francisco.compbio.ucsf.edu/-jacobson/plop_manual/plop_overview.html</a></td>
<td>Extensive conformation sampling, OPLS energy, sufficient energy minimization</td>
</tr>
<tr>
<td>COILS</td>
<td><a href="http://www.ch.embnet.org/software/COILS_form.html">http://www.ch.embnet.org/software/COILS_form.html</a></td>
<td>Scan database of known loops from PDB</td>
</tr>
<tr>
<td>MODELLER (loop module)</td>
<td><a href="http://guitar.rockefeller.edu/modeller/modeller.html">http://guitar.rockefeller.edu/modeller/modeller.html</a></td>
<td>Ab-initio conformation sampling plus CHARMM force fields</td>
</tr>
</tbody>
</table>
3.3.6 Side-chain modeling

The greatest success in the prediction of side-chain conformations has been achieved for core residues where packing constraints significantly simplify the problem. Even for core residues, the accuracy of side-chain prediction degrades when the structure of the backbone is itself not known to a high degree of accuracy. Many side-chain programs are based on rotamer libraries (Ponder et al., 1987), which are generally defined in terms of side-chain torsional angles for preferred conformations of a particular side chain. The resolution of rotamer libraries has increased over time and rotamer libraries have been compiled simply by sampling all angles at some given level of resolution. Since backbone conformation changes the frequency of the rotamers, backbone-dependent rotamer library is often used in side chain modeling. The major advantage is to increase computing efficiency, since bad rotamers, e.g. clashing with the backbone, have been automatically removed during construction of the rotamer library. Baker and his coworkers have developed a "solvated rotamer" approach that shows improvement on side chain packing at protein-protein interface. This approach extends current side-chain packing methods by using a rotamer library including solvated rotamers with one or more water molecules fixed to polar functional groups in probable hydrogen bond orientations, together with a simple energetic description of water-mediated hydrogen bonds. As the number of rotamers increases, however, so does the problem of sampling all possible conformations. There have been a variety of approaches developed to deal with the combinatorial problem in side-chain prediction. In recent papers, accuracies of about 1Å rmsd have been reported for core residues in known structures where the backbone has been fixed in the native conformation. A number of recent studies suggest that further improvements may still be possible. Mendes et al., (1999) found, for example, that the use of an intrinsic torsional potential can improve prediction accuracy. Recently reported a novel rotamer library in which
internal clashes between side chain and backbone are removed. This library
could, in principle, be used to improve prediction accuracy. We have recently
shown that using a very detailed rotamer library, which is based on rotamers that
use Cartesian coordinates taken from known structures rather than idealized bond
lengths and angles, yields rmsd values relative to the native of only 0.62Å for core
residues (Xiang et al., 2001). This appears to constitute a significant improvement
over existing procedures and demonstrates that the combinatorial problem,
usually assumed to greatly complicate side-chain prediction, may in fact be of little
consequence. This was later confirmed in more detailed study (Desmet et al.,
2002), which showed that low order local minima for side chain prediction may be
almost as accurate as the global minimum when evaluated against experimentally
determined structures. Improvement on side chain prediction in recent years has
mainly come from better energy functions. Eyal et al., (2004) showed that solvent
accessibility and contact surface area are important on the accuracy of side chain
prediction, particularly for modeling buried side chains. Liang and Grishin (2002)
have developed a new and simple scoring function for side chain prediction that
consists of the following energy terms: contact surface, volume overlap, backbone
dependency, electrostatic interactions, and desolvation energy. The weights of
these energy terms were optimized to achieve the minimal average root mean
square (rms) deviation between the lowest energy rotamer and real side-chain
conformation on a training set of high-resolution protein structures. The derived
scoring function combined with a Monte Carlo search algorithm was used to place
all side chains onto a protein backbone simultaneously. The average prediction
accuracy was 87.8% for chi (1), 73.2% for chi (1 + 2), and 1.34 Angstrom rms
deviation for all side chains in a protein structure. As is the case for loop
prediction, side-chain prediction accuracy depends sensitively on the accuracy to
which the backbone conformation is known (Huang et al., 1998). This suggests
the possibility of developing procedures where side-chain and backbone
conformation can be used iteratively to refine homology models.
Table 3.3 lists some publicly available side chain prediction programs and the
methods they used. Earlier side chain predictions, e.g. RAMP, Confmat etc.,
were usually based on small rotamer libraries; most recent programs were using
very detailed rotamer libraries, e.g., SCAP SCWRL SMOL (Liang et al., 2002).
### Table 3.3 Side Chain Modeling Programs

<table>
<thead>
<tr>
<th>Programs</th>
<th>Availability</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAP</td>
<td><a href="http://trantor.bioc.columbia.edu/programs/jackal/">http://trantor.bioc.columbia.edu/programs/jackal/</a></td>
<td>Colony energy methods with simple energy and large Cartesian-coordinate rotamer library</td>
</tr>
<tr>
<td>SCWRL</td>
<td><a href="http://dunbrack.fccc.edu/SCWRL3.php">http://dunbrack.fccc.edu/SCWRL3.php</a></td>
<td>Simple energy with back bone dependent rotamer library</td>
</tr>
<tr>
<td>SCCOMP</td>
<td><a href="http://atlantis.weizmann.ac.il/-eyale/">http://atlantis.weizmann.ac.il/-eyale/</a></td>
<td>Optimized scoring function and Gibbs sampling like algorithm</td>
</tr>
<tr>
<td>RAMP</td>
<td><a href="http://www.ram.org/computing/ramp/ramp.html">http://www.ram.org/computing/ramp/ramp.html</a></td>
<td>Knowledge based potentials and small rotamer library</td>
</tr>
<tr>
<td>SMD</td>
<td><a href="http://condor.urbb.jussieu.fr/smd.php">http://condor.urbb.jussieu.fr/smd.php</a></td>
<td>Flex force field, small rotamer library and dynamic cluster analysis of known structures</td>
</tr>
<tr>
<td>Confimat</td>
<td>Contact Koehl at <a href="mailto:Koehl@csb.stanford.edu">Koehl@csb.stanford.edu</a></td>
<td>Self-consistent mean field and small rotamer library</td>
</tr>
<tr>
<td>Maxsprut</td>
<td><a href="http://www.ebi.ac.uk/maxsprut/">http://www.ebi.ac.uk/maxsprut/</a></td>
<td>Rough energy function and small rotamer library</td>
</tr>
</tbody>
</table>

For partially buried residues, SMOL performed the best, which was due to its more sufficient conformation sampling and optimized scoring function. SCWRL performed reasonably well though not as accurate as the other two, but with much less CPU cost. On a 300MH SGI machine, SCWRL is very fast, 3 seconds for each protein, while SMOL needs 11700 seconds and SCAP needs 361 seconds. In the recent CASP6, SCWRL was ranked the first in the category of side chain.
3.3.7 Homology model refinement

High-resolution refinement is a difficult task that requires an effective sampling strategy as well as an accurate energy function to guide the search through conformational space. Homology model refinement is primarily focused on tuning alignment and modeling loops and side chains. Loops are usually the most variable regions of a structure where insertion and deletion often occur. Correct alignment is the most important task for homology modeling since the errors introduced into the model by misalignment are hard to remove in the later stages of refinement. When the sequence identity is above 40%, errors in the homology structure mainly come from side chains; when the sequence identity is between 30-40%, loops and side chains become most problematic (Sanchez et al., 1997). Given a good energy function, loop and side-chain refinement can in principle be applied repeatedly to relax the backbone closer to the native. Refinement on helix and beta sheet can be handled with similar methods as for loops, where proper hydrogen bond constraints should be applied to retain the secondary structure definition. Recent attempts have been made to use physical chemistry energy to refine side chains, loop and helix segments, target-template alignment, and the whole model.

3.3.8 Model assessment

All models built by homology will have errors as discussed in the previous section. Verification of the model and estimation of the likelihood and magnitude of errors has become one of the most important steps in advancing the state of the art of homology modeling. Errors of the model are usually estimated either from the energy of the model or from the resemblance of a given characteristic of the model to real structures. The most critical component is the development of a scoring function that is capable of discriminating good and bad models, and therefore, would have enormous impact on the ability to predict protein conformations. Scoring functions used for the evaluation of protein models generally fall into two broad categories. 'Statistical' effective energy functions (Sippl, 1995) are based on the observed properties of amino acids in known structures, and have been widely used in fold recognition and homology modeling applications. A variety of statistical criteria have been used successfully to discriminate
between deliberately misfolded and native structures. Most of them are directly or indirectly based on the analysis of contacts, either inter residue contacts, inter atom contacts, or contacts with solvent. For example, preferential distributions of polar and apolar residues inside or outside of protein can be used to detect completely misfolded, salvation potentials can detect local errors as well as complete misfolds, packing rules have been implemented for structure evaluation. Residue or atom contacts are discriminative because they are energetically favored. Real structures cannot tolerate too much unfavorable interaction. Thus for a model to be correct only a few infrequently observed atomic contacts are allowed. However, bond angles and bond lengths, though powerful in checking the quality of experimental structures, are usually less useful for the evaluation of models because these factors have already been considered appropriately in the model building stage (Fiser et al., 2003). Although they are computational cheap, statistical energies are not sensitive to evaluate near native decoy structures, especially for segments of proteins, such as in the modeling of loop and side chains. Physical effective energy functions are based on a direct evaluation of the conformational free energy of a protein. Recent work has demonstrated that such a direct evaluation of the conformational free energy can be at least as successful as statistically based scoring functions in distinguishing the native structure of a protein from an incorrectly folded decoy, although generally at greater computational. A distinct advantage of such physically derived functions is that they are based on well-defined physical interactions, thus making it easier to learn and to gain insight from their performance. Moreover, the success in CASP of ab-initio methods based on purely physical chemistry methods suggests that our understanding of the forces that drive protein stability may have reached the point where it can be translated into widely applicable computational tools. One of the major drawbacks of accurate physical chemical description of the folding free energy of a protein is that the treatment of solvation required usually comes at a significant computational expense. Fast solvation models such as the Generalized Born and a variety of simplified scoring schemes (Petrey et al., 2000) may prove to be extremely useful in this regard.

A number of freely available programs can be used to verify homology models as shown in Table 3.4. They generally belong to one of two categories. The first category (e.g. PROCHECK and WHATIF) checks for proper protein
stereochemistry, such as symmetry checks, geometry checks (chirality, bond lengths, bond angles, torsion angles, etc) and structural packing quality.

Table 3.4 Model assessment programs

<table>
<thead>
<tr>
<th>Programs</th>
<th>Availability</th>
<th>Quality to be checked</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROCHECK</td>
<td><a href="http://www.biochem.ucl.ac.uk/-roman/procheck/procheck.html">http://www.biochem.ucl.ac.uk/-roman/procheck/procheck.html</a></td>
<td>Stereochemistry</td>
</tr>
<tr>
<td>WHATCHECK</td>
<td><a href="http://www.sander.embl-heidelberg.de/whatcheck/">http://www.sander.embl-heidelberg.de/whatcheck/</a></td>
<td>Stereochemistry, nomenclature, symmetry, packing, inside/outside profile, missing atoms/residues, hydrogen bonds, etc</td>
</tr>
<tr>
<td>ProsA II</td>
<td><a href="http://www.came.sbg.ac.at">http://www.came.sbg.ac.at</a></td>
<td>Mis-folded structures, faulty parts of structural models</td>
</tr>
<tr>
<td>VERIFY3D</td>
<td><a href="http://www.doe-mbi.ucla.edu/services/Verify-3D/">http://www.doe-mbi.ucla.edu/services/Verify-3D/</a></td>
<td>Residues fitness in the model environment</td>
</tr>
<tr>
<td>ERRAT</td>
<td><a href="http://www.doe-mbi.ucla.edu/services/Errat.html">http://www.doe-mbi.ucla.edu/services/Errat.html</a></td>
<td>Statistical non-bonded atom-atom interactions</td>
</tr>
<tr>
<td>ANOLEA</td>
<td><a href="http://www.fundp.ac.be/pub/ANOLEA.html">http://www.fundp.ac.be/pub/ANOLEA.html</a></td>
<td>Non-local environment of heavy atoms</td>
</tr>
<tr>
<td>AQUA</td>
<td><a href="http://www.nmr.chem.uu.nl/users/jurgen/Aqua/server/">http://www.nmr.chem.uu.nl/users/jurgen/Aqua/server/</a></td>
<td>Violation, completeness and redundancy of NOE distance restraints</td>
</tr>
<tr>
<td>Programs</td>
<td>Availability</td>
<td>Quality to be checked</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------</td>
<td>------------------------------------------------------------</td>
</tr>
<tr>
<td>Probe</td>
<td><a href="http://kinemage.biochem.duke.edu/software/probe.pdb">http://kinemage.biochem.duke.edu/software/probe.pdb</a></td>
<td>Atomic packing, either within or between molecules</td>
</tr>
<tr>
<td>SQUID</td>
<td><a href="http://www.ysbl.york.ac.uk/oldfield/squid/">http://www.ysbl.york.ac.uk/oldfield/squid/</a></td>
<td>Analysis and display of data from crystallography and molecular dynamics</td>
</tr>
<tr>
<td>PROVE</td>
<td><a href="http://www.ucmb.ub.ac.be/UCMB/PROVE">http://www.ucmb.ub.ac.be/UCMB/PROVE</a></td>
<td>Check the quality of the atomic model of a macromolecule structure based on the calculations of atomic volumes</td>
</tr>
<tr>
<td>GRASP2</td>
<td><a href="http://tranter.bioc.colombia.edu/programs.html">http://tranter.bioc.colombia.edu/programs.html</a></td>
<td>Graphic display model structure and sequence-template alignment</td>
</tr>
</tbody>
</table>

The second category (e.g., VERIFY3D and PROSAII) checks the fitness of sequence to structure, and assigns a score for each residue fitting its current environment. A new graphics software called GRASP2 developed in the Honig lab is also extremely useful in model assessment (Petrey et al., 2003). The software can display alignments and template structures simultaneously for assessment of the alignment quality. For example, gaps and insertions can be mapped to the structures to verify that they make sense geometrically. Where residue substitutions occur, the user can verify that structural features such as hydrophobic packing are maintained and that active-site residues and other features of the target identified from the literature are conserved. The manual inspection should be combined with existing programs to further identify problems in the model.
3.4 Application of homology models

There are numerous applications for protein structure information hence homology models are used at various stages of the drug discovery process. The most spectacular success is to identify or to optimize compounds that were subsequently progressed to clinical trials or to the drug market. It can still be helpful in proposing and testing hypothesis in molecular biology, such as hypotheses about the location of ligand binding sites, substrate specificity, function annotation, and drug design (Takeda et al., 2004). It can also provide starting models for solving structures from X-ray crystallography, NMR and electron microscopy. Homology models are used in the rapid detection, description and analysis of ligand-binding pockets, in predicting antigenic epitopes (Sali and Blundell, 1993), in simulating protein-protein docking, in inferring function from calculated electrostatic potential around the protein, in facilitating molecular replacement in X-ray structure determination, in refining models based on NMR constraints, in testing and improving a sequence-structure alignment (Wolf et al., 1998), in annotating single nucleotide polymorphisms, in structural characterization of large complexes by docking to low-resolution cryo-electron density maps; and in rationalizing known experimental observations. The design of site-directed mutant proteins is one further important option for the application of homology models to target validation. Introducing point mutations and
subsequently studying the effects in vitro or in vivo is a common approach in molecular biology. Site-directed mutants of the target protein can be made to render that target sensitive to an existing pharmacological agent. Based on homology models, some members of the mitogen-activated protein (MAP) kinase family were mutated to make them sensitive to a kinase inhibitor from the pyridinyl imidazole class. This enabled the use of the compound for broader target validation studies. There are several examples where protein homology models have supported the discovery and the optimization of lead compounds with respect to potency and selectivity. Currently, the structures of 40 of the 518 known different human protein kinases have been characterized by X-ray crystallography (Manning, 2002). Homology model-based drug design has been applied to epidermal growth factor-receptor tyrosine kinase protein, Bruton’s tyrosine kinase, Janus kinase 3 and human aurora 1 and 2 kinases. Using the X-ray crystal structure of cyclin-dependent kinase 2 (CDK2), generated a homology model of CDK4. This model guided the design of highly potent and selective CDK4 inhibitors that were targeted towards the ATP binding pocket Siedlecki et al., (2003) have demonstrated the utility of homology modeling in the prediction of pharmacologically active compounds. Alterations in DNA methylation patterns play an important role in tumorigenesis; therefore, inhibitors of DNA methyltransferase 1 (DNMT1), which is the protein that represents the major DNA methyltransferase activity in human cells, are desired. Known inhibitors from the 5-azacytidine class were docked into the active site of a DNMT1 homology model, which led to the design of N4-fluoroacetyl-5-azacytidine derivatives that acted as highly potent inhibitors of DNA methylation in vitro. Thrombin-activatable fibrinolysis inhibitor (TAFI) is an important regulator of fibrinolysis, and inhibitors of this enzyme have potential use in antithrombotic and thrombolytic therapy: Based on a homology model of TAFI (~50% sequence identity to carboxypeptidases A and B), appropriately substituted imidazole acetic acids were designed and were subsequently found to be potent and selective inhibitors of activated TAFI (Barrow et al., 2003), appropriately substituted imidazole acetic acids were designed and were subsequently found to be potent and selective inhibitors of activated TAFI (Barrow et al., 2003).
Fig 3.7 Accuracy and application of protein structure models. Different ranges of applicability of comparative protein structure modeling, threading, and denovo prediction; the corresponding accuracy of protein models; and their sample applications are shown (Baker and Sali, 2001).

Plenty of examples for the successful application of homology modeling in drug discovery are described above. In the absence of experimental structures of drug target proteins, homology models have supported the design of several potent pharmacological agents. One of the advantages of homology models is that these models can be generated relatively easily and quickly. Furthermore, such models could support the hypotheses of medicinal chemists on how to generate biologically active compounds in the important early conceptual phase of a drug discovery project. The design of compounds that are selectively directed at particular drug target proteins is one of the strengths of this technique. Such selective compounds can even be applied to gain insights into the physiological role of novel drug targets. However, while complete experimental structures of pharmacologically important proteins are missing; the homology modeling technique provides one approach to bridge the gap until this information becomes available.
3.5 Introduction to Molecular dynamics

Molecular dynamics (MD) simulations are important tools for understanding the physical basis of the structure and function of biological macromolecules. The early view of proteins as relatively rigid structures has been replaced by a dynamic model in which the internal motions and resulting conformational changes play an essential role in their function. Simulations can provide the ultimate detail concerning individual particle motions as a function of time. Thus they can be used to address specific questions about the properties of a model system, often more easily than experiments on the actual system. Of course, experiments play an essential role in validating the simulation methodology: comparison of simulation and experimental data serve to test the accuracy of the calculated results and to provide criteria for improving methodology. This is particularly important because theoretical estimates of systematic errors inherent in simulations have not been possible—that is, the errors introduced by the use of empirical potentials are difficult to quantify. Another significant aspect of simulation is that, although the potentials used in simulations are approximate, they are completely under the control of the user, so that by removing or altering specific contributions their role in determining a given property can be examined. This is most graphically demonstrated by the use of "computer alchemy"—transmuting the potential from that representing one system to another during a simulation in the calculation of free energy differences" (Simonson, 2002).

There are three types of applications of simulation methods in the macromolecular area, as well as in other areas involving mesoscopic systems. The first uses simulation simply as a means of sampling configuration space. This is involved in the utilization of MD, often with simulated annealing protocols, to determine or refine structures with data obtained from experiments. The second uses simulations to obtain a description of the system at equilibrium, including structural and motional properties (for example, atomic-square fluctuation amplitudes) and the values of thermodynamic parameters. For such applications, it is necessary that the simulations adequately sample configuration space, as in the first application, with the additional condition that each point be weighted by the appropriate Boltzmann factor. The third area uses simulations to examine the actual dynamics. Here not only is adequate sampling of configuration space with the appropriate Boltzmann
weighting required, but it must be done so as to correctly represent the development of the system over time. For the first two areas, Monte Carlo simulations can be used, as well as MD. By contrast, in the third area where the motions and their development with the time are of primary interest, only MD can provide the necessary information. The three sets of applications make increasing demands on simulation methods as to their required accuracy and precision.

3.5.1 Molecular dynamics today
The increase in the number of studies using MD to stimulate the properties of biological macromolecules has been fueled by the general availability of programs and of the computing power required for meaningful studies. The original simulation was less than 10 ps in length. By comparison, current simulations are often 1,000 times as long (10 ns) but take a factor of ~50 less time for a system of same size. Much of the gain in simulation time is re-invested into studying much larger systems ($10^4$-$10^6$ atoms instead of 500) to include, for example an explicit solvent and/or membrane environment. In addition to being able to perform simulations for a much longer period to time, another important consequence of the access to faster computers is that multiple simulation runs can be performed to obtain estimates of statistical errors. Although the widely used programs (such as CHARMM (Brooks, 1983) and its direct descendants AMBER (Weiner, 1981) and GROMACS (Scott, 1999) have a great range of capabilities, users often find that they can not do exactly what is needed to solve their particular problem. Thus, developments in simulation methodology continue apace with applications (Tuckerman, 2000). One of the very exciting recent developments in MD simulations is that, with modern computers, the simulation time is extended to a range from 100 ns to μs, making it possible to study biological phenomena as they happen. This is analogous but in an inverse sense to the observation that, although experiments on the ps time scale were an important development, it is only when the time resolution was extended to femto seconds that the actual events involved in chemical reactions could be observed (Polanyi, 1995). A striking recent result is that, by running multiple simulation of 10 ns duration, the "real time" visualization of water molecule migrating through a model of the aquaporin channel has been achieved. It is becoming increasingly evident that essential functions in the cell are executed not by individual proteins, but by
protein complexes. The structures of such large subunit complexes are now being determined. In most complexes conformational change is directly involved in function. Two such systems that have already been simulated by MD are the nicotinic acetylcholine receptor and ATP synthase (Bockmann, 2002). With the pace of advance in both computer hardware and algorithms, simulation of such larger systems for the time required to obtain meaningful results is expected to be possible in the very near future.

The next stage for MD simulations is the evolution from molecular and supramolecular systems to the cellular level. Studies of the formation of such assemblages will certainly be more demanding. An interesting example of structure formation studied in real time by MD concerns the formation of phospholipid bilayers. The simulation of more complex cellular activities, such as synaptic transmission and the dismantling of the nuclear membrane on cell division by the cytoplasmic motor protein dynein are two examples of interest. Much of this work will build on the detailed knowledge of the structure and dynamics of channels, enzymes and other cellular components. MD simulations of biological macromolecules have provided many insights concerning the internal motions of these systems since the first protein was studied over 25 years ago. With continuing advances in the methodology and the speed of computers, which is still doubling every eighteen months, MD studies are being extended to larger systems, greater conformational changes and longer time scales. This makes possible the investigation of motions that have particular functional implication and to obtain information that is not accessible from experiment. The results available today make clear that the applications of MD will play an even more important role for our understanding of biology in the future. Computation provides an excellent method to identify the individual conformational states of biological systems and their transitions between different conformational states. Both can be predicted using MD. It calculates the time dependent behavior of a molecular system. It was introduced by Alder and Wainwright in 1957. MD simulations provide detailed information on the fluctuations and conformational changes of proteins and nucleic acids, and they are now routinely used to investigate the structure, dynamics and thermodynamics of biological molecules and their complexes. MD may have a role in the improvement of the medium resolution (approximately 4 Å) structures and homology models. And indeed, as well as restricted MD being
used in some solved structure refinement programs, it has also been used in the 
completion/refinement of homology models (Flohil et al., 2002). In MD, forces on 
atoms, due to interactions with other atoms, are computed using certain empirical 
force fields. Once force can be computed, Newton’s laws of motion are used, 
almost always with an explicit time integration scheme, to determine the trajectory 
of the system. The force fields, in turn, are pre-determined by approximating the 
results of quantum mechanical calculations and experiments on small protein 
fragments. Quantum mechanical calculations can be used to directly determine 
forces for use in MD, rather than to determine an empirical force field, but current 
quantum mechanical methods are too slow to be feasible (Clore et al., 1986). The 
objectives of MD simulations are two-fold: (i) to determine a statistically 
representative set of conformational states and (ii) to reproduce the dynamical 
transitions between these states. Monte Carlo methods can, in principle, be used 
to determine such ensembles of states. However, they have shown poor 
efficiency for large, dense systems compared with molecular dynamics. MD 
simulations are mainly based on Newton’s second law of motion, formulated in the 
equation \( F = ma \), with \( F \) the forces exerted on a particle, \( m \) it’s mass, and its 
acceleration. The simulations revolve around solving Newton’s equation of motion 
for a system of \( N \) interacting atoms (Hess et al., 2008).

\[
F_i = m_i \frac{\partial^2 r_i}{\partial t^2}, \quad i = 1, ..., N \tag{2.1}
\]

where the forces \( F_i \) are the negative derivatives of a potential function \( V (r_1, \ r_2, ..., r_N) \):

\[
F_i = -\frac{\partial V}{\partial r_i} \tag{2.2}
\]

At each step of the simulation, the forces and the positions of the atoms are 
calculated using Eqs. (2.1) and (2.2), and written to file as output, so the 
trajectories of the atoms through system-space can be viewed later. Energy 
minimization (EM) is the act of finding a minimum in the potential energy 
landscape of a molecular system. In the system there exists by definition one 
global minimum, and a large number of local minima. Mapping all of these 
minima is almost impossible however, and thus we will be looking only for the
nearest local minimum. A way to do this is by using the steepest-descent method, which is easy to implement, is guaranteed to find the minimum within a reasonable number of steps, and each iteration is fast. It does not converge very fast however when host get closer to the minimum.

### 3.5.2 Limitation on Time Scale

A limitation of MD is the short time scale that can be accessed via simulation. Large-scale protein conformational changes, such as folding and allosteric transitions, normally don't occur in under a millisecond. MD on proteins, limited by high frequency motions to time steps of about a femto second \(10^{15}s\) (Fig. 3.8), can currently access only about a micro- second of real time. In fact, this limitation of MD has been identified as one of important challenges in computational biology (Computational science, 2005).

Fig. 3.8: Typical motional timescales for physical processes.

To illustrate this problem, consider a 10000-atom simulation with time step size 1 femto second, carried out to a millisecond of simulation time. Each iteration will require approximately 50 ms of wall-clock time with a fast code, such as GROMACS. The \(10^{12}\) iterations will then require \(5 \times 10^{10}\) s, which is around 1600 years of sequential computing time. Using massive parallelism, on say 32000 processors, we can solve the same problem in eighteen days, if we obtain high efficiency.
3.5.3 CHARMM
The CHARMM program allows generation and analysis of a wide range of molecular simulations. The most basic kinds of simulation are minimization of a given structure and production runs of a MD trajectory. More advanced features include free energy perturbation (FEP), quasi harmonic entropy estimation, correlation analysis and combined quantum, and molecular mechanics (QM/MM) methods. CHARMM is one of the oldest programs for MD. It has accumulated a huge number of features, some of which are duplicated under several keywords with slight variations. This is an inevitable result of the large number of outlooks and groups working on CHARMM throughout the world. The change log file as well as CHARMM's source code is good places to look for the names and affiliations of the main developers. The involvement and coordination by Charles L. Brooks III's group at the University of Michigan is very salient (Brooks et al., 1983; MacKerell et al., 1998b).

3.5.4. Gromacs
Gromacs (Groningen Machine for Chemical Simulations) is a MD simulation package originally developed (Hess et al., 2008) at the Department of Biophysical Chemistry of Groningen University, now maintained and extended at different places. The Gromacs project was originally started to construct a dedicated parallel computer system for molecular simulations, based on ring architecture. The MD specific routines were rewritten in the C programming language from the Fortran77-based program Gromocs, which had been developed in the same group. The program is written for Unix-like operating systems; it can run on Windows machines if the Cygwin Unix layer is used. The program can be run in parallel on multiple CPU cores or a network of machines using the MPI library. Gromacs contains a script to convert molecular coordinates from a PDB file into the formats it uses internally. Once a configuration file for the simulation of several molecules (possibly including solvent) has been created, the actual simulation run (which can be time consuming) produces a trajectory file, describing the movements of the atoms over time. This trajectory file can then be analyzed or visualized with a number of supplied tools. The concept used in Gromacs is periodic boundary condition and group. Periodic boundary condition is classical way used in Gromacs to reduce edge effect in system. The atom will be placed in
a box, surrounded by a copy of the atom (Fig. 3.9). In Gromacs there are some model boxes. That is triclinic, cubic, and octahedron. The second concept is group. This concept is used in Gromacs to show an action. Each group can only have a maximum number of 256 atoms, where each atom can only have six different groups (Berendsen et al., 1995).

THE GLOBAL MD ALGORITHM

1. Input initial conditions
Potential interaction V as a function of atom positions
Positions r of all atoms in the system
Velocities v of all atoms in the system

repeat 2,3,4 for the required number of steps:

2. Compute forces
The force on any atom $F_i$ is computed by calculating the force between non-bonded atom pairs:
$F_i = \sum_{j \neq i} F_{ij}$
plus the forces due to bonded interactions (which may depend on 1, 2, 3, or 4 atoms), plus restraining and/or external forces.
The potential and kinetic energies and the pressure tensor are computed.

3. Update configuration
The movement of the atoms is simulated by numerically solving Newton's equations of motion
$\frac{d^2}{dt^2} r_i = F_i / m_i$
$\frac{dr_i}{dt} = v_i$
$\frac{dv_i}{dt} = F_i / m_i$

4. if required Output step
write positions, velocities, energies, temperature, pressure, etc.

Fig 3.9 Gromacs steps for performing MD simulations.

3.5.5 Introduction to Molecular Docking
Molecular docking predicts the orientation of the ligands bound to receptors by assuming that the receptor conformation is known (Taylor et al., 2002). Molecular docking is a fast method to explore substrate/receptor complexes in the field of drug discovery as well as in understanding biochemical processes. The major techniques used for molecular docking are the following, Genetic algorithms,
Simulated annealing, Molecular dynamics, Monte Carlo methods, Distance
geometry methods, Point complementary methods, Fragment-based methods,
Tabu searches and Systematic searches. Docking procedures are composed of
two components of which the first is a search algorithm and later is a scoring
function. The search algorithm finds different conformations for the ligand by using
one of the methods listed above. Systematic searches explore all possible binding
modes between the ligand and receptor. However, this takes a huge amount of
computational time especially for large flexible ligands. The amount of
conformational space explored and the computational time required for the search
must be balanced. Scoring functions are used to rank the different conformations
obtained by the search algorithm. A good scoring function distinguishes the
experimentally obtained conformation from all other conformations explored
through the search algorithm. The major scoring methods are empirical free
energy scoring functions, molecular mechanics force fields, and knowledge-based
functions (Taylor et al., 2002). Some of the docking programs currently in use are
the DOCK (Kuntz et al., 1992), GOLD (Jones et al., 1997), AutoDock (Morris et
al., 1998), Surflex (Jain et al., 1998), FlexX, FTDOCK. The differences between
them are derived from the different search algorithms or different implementation
of the same algorithms, and different scoring functions. Most of these programs
hold the receptor rigid and allow a certain degree of flexibility to the ligands.

3.5.6 Genetic algorithms

The Genetic algorithms (Judson et al., 1997) are considered to be stochastic
global optimization methods. Since they do not use gradient information, they can
be applied various different types of optimization problems. The genetic algorithm
searches the parameters forming the conformational space simultaneously.
Judson (Judson et al., 1997) simply explains the genetic algorithms by using
Figure 3.10 as an example.
Three local minima labeled as I, II (the global minimum), and III are shown in the f(x) function. Then, the genetic algorithm terms are introduced: fitness, populations, and chromosomes. The function f(x) is the fitness. The populations are set of individuals sampling the conformational space. The x in the example denotes for one of the chromosomes, which are the parameters forming each individual. Other genetic algorithm terms are mutations, selection, crossovers, and migrations (Westhead et al., 1997). The mutation operator obtains the individuals by randomly changing the chromosomes. The best individuals based on the fitness function are then selected for crossover. The crossover allows an exchange of a set of chromosomes from one parent to another. The migration moves individual chromosomes from one sub-population to another. The mutation, crossover, and migration procedures continue until some stopping criteria are met. The following paragraphs give an introduction about three docking programs of which AutoDock is considered the most commonly used program and hence it has been used in the context of present thesis work also.

3.5.7 FlexX

In FlexX (Rarey et al., 1996a, 1996b, 1997), interaction types and geometries according to Bohm (Bohm, 1992a, b) and Klebe (Klebe, 1994) describe the protein ligand interactions. Each interacting group of the molecule to be docked is assigned an interaction type and a corresponding compatibility. Possible
interaction types are geometrically restricted hydrogen bonds, interactions between metals and metal acceptors, and hydrophobic interactions, for example those between phenyl rings and methyl groups. For each group capable to form an interaction, special contact geometry is defined by placing an interaction surface around the centre, usually as part of a sphere. Two groups form an interaction if the interaction centre of one group coincides with the interaction surface of a counter group. The docking algorithm in FlexX is based on an incremental construction strategy consisting of three phases: In the first phase (base selection), the base fragment of the ligand is selected which is then placed into the active site of the protein (base fragment placement). Finally, the ligand is reconstructed in an incremental fashion, starting from different placements of the base fragment (complex construction). Upon connecting additional fragments, new interactions are screened and the best partial solutions based on the ranking of a scoring function are hooked up until the ligand is completely constructed. The docking algorithm is relatively sensitive to the selection and placement of the base fragment. If the geometry of a fragment of a molecule to be docked is known (e.g., from a similar ligand crystallized in the target protein), a useful option is to place the referring fragment manually into the binding pocket via the mapref command. This reduces the run time of the docking procedure and increases the probability of predicting the correct binding mode of the ligand. Another possibility to include knowledge about protein-ligand interactions a priori, as user-defined constraint, into the docking process can be realised by FlexX-Pharm, an extended version of FlexX. The constraints are determined by selected FlexX interactions and inclusion volumes. They guide the docking process to produce a set of docking solutions with particular properties. By applying a series of look-ahead checks during the flexible construction of ligand fragments within the active site, FlexX-Pharm determines which partially built docking solutions can potentially obey the constraints. Solutions that will not obey the constraints are discarded as early as possible, thus decreasing the calculation time and enabling new docking solutions to emerge.

3.5.8 AutoDock

The program AutoDock (Goodsell et al., 1998) was developed to provide an automated procedure for predicting the interaction of ligands with biomacromolecular targets. The docking simulation is carried out using one of a
number of possible search methods. The Lamarckian genetic algorithm and the Monte Carlo simulated annealing are available in AutoDock. The receptor is considered as rigid, and the allowed flexibility for the ligand is 52 torsion angles. The docking methodology in AutoDock is examined in the following four subsections: Preparation of the ligand and receptor, the Autotors and AutoGrid procedure, the genetic algorithm implementation, and the fitness function and evaluation of the free energy.

3.5.8.1 Preparation of the ligand and receptor
First the docking box has been placed into the active side of the receptor, a docking box with a grid consisting of 60 X 60 X 60 points (default parameters) and 0.575 Å grid spacing were employed in the present study. The box was oriented so that the long side was along the direction of the center of binding site to the center of the entrance of the binding site. In this orientation, the box included the entire binding site and some area just outside of the binding pocket entrance. The docking boxes of sizes between 82 X 60 X 60 and 110 X 80 X 80 with 4 grid point increments can also be used in the calculations. The size 92 X 70 X 70 has been found optimum so that the ligand is free to have different conformations in the binding cleft and also it is small enough to save the computational time.

3.5.8.2 AutoGrid Procedure
AutoDock requires pre-calculated grid maps, one for each atom type present in the ligand being docked. This helps to make the docking calculations extremely fast. These maps are calculated by AutoGrid. A grid map consists of a three dimensional lattice of regularly spaced points, surrounding (either entirely or partly) and centered on some region of interest of the macromolecule under study. Each point within the grid map stores the potential energy of a 'probe atom' or functional group that is due to all the atoms in the macromolecule. Figure 3.11 illustrates the main features of a grid map (Morris et al., 1998).
In the Figure 3.11 the whole protein is in the docking box which is defined by the grid points with user defined grid spacing. The energetics of a particular substrate configuration is found by tri-linear interpolation of affinity values of the eight grid points surrounding each of the atoms in the substrates. The time to perform an energy calculation using the grids is proportional only to the number of atoms in the substrate, and is independent of atoms in the protein.

3.5.8.3 Genetic Algorithm Implementation

The particular arrangement of a ligand and a protein is defined by state variables, which include a set of variables describing the translation, orientation, and conformation of the ligand with respect to protein. Each state variable corresponds to a gene, and the ligand's state corresponds to a genotype, whereas its atomic coordinates correspond to the phenotype. In AutoDock implementation (Morris et al., 1998), the chromosome is composed of real valued genes: three Cartesian coordinates for the ligand translation; four variables defining a quaternion specifying the ligand orientation; and one real-value for each ligand torsion angle, in that order. The genetic algorithm begins by creating a random population of individuals, where the number of individuals in the population is user defined. For each random individual in the population, random values are assigned for the genes. A loop over generations then takes place,
repeating until the maximum number of generations or the maximum number of energy evaluations is reached, whichever comes first. A generation consists of five stages: mapping and fitness evaluation, selection, crossover, mutation, and elitist selection, in that order. Mapping translates from each individual's genotype to its corresponding phenotype. This allows each individual's fitness to be evaluated. The fitness function and the energy evaluation are explained in the next sub-section. Every time an individual's energy is calculated. This is followed by proportional selection to decide which individuals will reproduce. Crossover and mutation are performed on random members of the population according to user-defined rates of crossover and mutation. First, crossover is performed. The new members are replaced the parents in the population, keeping the population size constant. Crossover is followed by mutation. Optional user-defined integer parameter elitism determines how many of the top individuals also automatically survive into the next generation. The genetic algorithm iterates over generations until one of the termination criteria is met.

3.5.8.4 The Fitness Function and Free Energy Calculation

The fitness is the sum of the intermolecular interaction energy between the ligand and the protein, and the intramolecular energy of the ligand. At the end of each docking, AutoDock reports the fitness (the docked energy), the state variables, the coordinates of the docked conformation, and the estimated free energy of binding $\Delta G$.

$$
\Delta G = \Delta G_{\text{vdw}} \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + \Delta G_{\text{el.}} E(t) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^6} \right) + \Delta G_{\text{elec}} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij})} + \Delta G_{\text{tor}} N_{\text{tor}} + \Delta G_{\text{sol}} \sum_{i,j} \left( S_j V_j + S_j V_i \right) e^{(r_{ij}^2/\sigma^2)}
$$

(2)

$\Delta G$ calculation including the solvation effect in AutoDock is shown in Equation 2.

The Lennard-Jones 12-6 dispersion-repulsion term is first, the second is $a_t$, based on the angle $t$, between the probe and the target atom, the third term is a screened Coulombic electrostatic potential, the fourth term is a measure of the unfavorable entropy of ligand binding due to the restriction of conformational
degrees of freedom, which is proportional to the number of sp5 bonds in the ligand, Ntor and finally the last term accounts for the desolvation effect. For desolvation, the pair wise, volume-based method of Stouten et al. is used. This method has the advantage that it is consistent with the pre-calculated affinity grid formulation used by AutoDock (Morris et al., 1998). For each atom in the ligand, fragmental volumes of surrounding protein are weighted by an exponential function and then summed, evaluating the percentage of volume around the ligand atom that is occupied by protein atoms (Morris et al., 1998). This percentage is then weighted by the atomic solvation parameter of the ligand atom to give the desolvation energy (Morris et al., 1998). The full method may be broken into four separate components: burial of polar atoms in the ligand, burial of a polar protein atom, burial of polar and charged atoms in the ligand, and burial of polar and charged protein atoms (Morris et al., 1998). Great success has also been reported in using simply the amount of hydrophobic surface area buried upon complexation as a measure of the 'hydrophobic effect', so they tested several formulations that included only the volume lost around ligand carbon atoms (Morris et al., 1998). The burial of polar atoms caused particular problems. Apart from the volume-based method, a simpler formulation for the solvent transfer of polar atoms has been tested, where a constant term corresponding to the favorable free energy of interaction of a polar atom with solvent is estimated, and this is subtracted from the binding free energy (Morris et al., 1998).

3.5.9 GOLD
GOLD (Jones et al., 1997) (Genetic Optimization for Ligand Docking) is an automated docking program that uses genetic algorithm to search the conformational space and allows full flexibility for acyclic ligands and partial protein flexibility in the neighborhood of the flexibility for acyclic ligands and partial protein flexibility in the neighborhood of the protein binding cleft (Jones et al., 1997). The docking methodology of GOLD will be examined in three parts: initialization of the protein and ligand, genetic algorithm implementation, and the fitness function.

3.5.9.1 Initialization of the protein and ligand
GOLD needs a user docking sphere that is placed in the binding cleft of the protein with a user defined center and a radius. The center can be a point, or an atom. Since there are X-ray crystal structures for ligands bound to HLA-A2.1
receptors, the binding site of the receptor is already known. The center of the binding cleft has been placed by using CHARMM program (Brooks et al., 1983). The receptor atoms within the 10 Å distances of ligand atoms were selected and the center of the selected volume obtained by using 'stats' option. The coordinates of the center is then found (4.00, 16.1, -6.70) in x, y, z directions, respectively. As a radius going along with this center, 20, 25, and 50 Å have been used. The 20 Å radius was found optimum so that the computational time and the accuracy can be balanced. Fig 3.12 shows the docking sphere used in the calculations. The protein is considered as rigid except OH groups of SER, THR and TYR, and NH4+ group of LYS in the active site neighborhood. The ligand can be prepared as fully flexible. The simplest constraints to apply to the ligand are keeping the ring corners, amide bonds, planar nitrogens, and/or internal hydrogen bonds constant. The other constraints available are the covalent constraints, distance constraints, H-bond constraints, structure-based constraints, and similarity constraints. The default value for the number of runs, which varies between 1 and 50, is 10. The protein and ligand input files must be in pdb or tripos mol2 format. The latter is suggested. If the pdb format is used, the program will assign the partial charges using a modified Tripos force field. Since mol2 files contain the partial charge information, different force fields can be used to prepare the mol2 files.

Fig 3.12: The size and location of docking sphere. The centre coordinates are (4.00, 16.1, -6.70) in X Y Z directions, respectively with 20 Å radius has been used in docking calculations in GOLD.
3.5.9.2 Genetic Algorithm Implementation

Gold uses a steady-state operator-based genetic algorithm to sample the conformational space and ligand binding modes (Jones et al., 1997). This genetic algorithm is illustrated in seven steps:

1. A set of reproduction operators (crossover, mutation, etc.) is chosen. Each operator is assigned a weight.
2. An initial population is randomly created and the fitnesses of its members determined.
3. An operator is chosen using roulette wheel selection based on operator weights.
4. The parents required by the operator are chosen using roulette wheel selection based on scaled fitness.
5. The operator is applied and child chromosomes are produced. Their fitness is evaluated.
6. If not already present in the population, the children replace the least fit members of the population.
7. If 100000 operators have been applied stop otherwise go to step three.

There are three operators used: mutation, crossover, and migration. The mutation operator creates the individuals in each population by randomly changing the rotatable bonds in the protein and ligand. Torsion angle values vary between -180° and 180° in step-sizes of 1.4°. The default values for the number of individuals in one population and number of populations are 100 and 5, respectively. The crossover operator provides the exchange of chromosomes between the individuals. The migration operator copies an individual from one population to another population. Operators were chosen using roulette-wheel selection based on operator weights. These weights were chosen so that crossover and mutation were applied with equal probability and migration was applied 5% of the time. After reaching the default value of number of operators, 100000, the algorithm terminates. After a conformation of the ligand and protein binding cleft generated, the ligand is placed within the active site using a least squares fitting procedure. Then, its fitness score, which is explained in the next sub-section, is evaluated.

3.5.9.3 The fitness function

Two different scoring functions are available in Gold: Chemscore and GoldScore.
Chemscore was derived empirically from a set of 82 protein-ligand complexes for which measured binding affinities were available. The ChemScore function was trained by regression against measured affinity data. The calculation of free energy of binding ($\Delta G_{\text{binding}}$) is shown in Equation 1.

$$\Delta G_{\text{binding}} = \Delta G_0 + \Delta G_{\text{hbond}} + \Delta G_{\text{metal}} + \Delta G_{\text{lipo}} + \Delta G_{\text{rot}} \quad (1)$$

Each component of this equation is the product of a term dependent on the magnitude of a particular physical contribution to free energy and a scale factor determined by regression. The final chem score value is then obtained by adding a clash penalty and internal torsion terms, which prevents close contacts in docking and poor internal conformations. Covalent and constraint scores may also be included.

ChemScore = $\Delta G_{\text{binding}} + P_{\text{clash}} + C_{\text{external}}P_{\text{internal}} + (C_{\text{covalent}}P_{\text{covalent}} + P_{\text{constraint}}) \quad (2)$

The Gold Score fitness function is composed of four components: protein-ligand hydrogen bond energy (external H-bond), protein-ligand vanderwaals (vdw) energy (external vdw), ligand internal vdw energy (internal vdw) and ligand torsional strain energy (internal torsion). Optionally, a fifth component, ligand intra molecular hydrogen bond energy (internal H-bond), may be added. Empirical parameters used in the fitness function (hydrogen bond energies, atom radii and polarizabilities, torsion potentials, hydrogen bond directionalities, etc.) are taken from GOLD parameter file which can be edited by the user. The external vdw score is multiplied by a factor of 1.575 when total fitness score is computed. This is an empirical correction to encourage protein-ligand hydrophobic contact. The final Gold score value is stated in the following expression.

Gold Score = $- (\text{H-Bond-Energy} + \text{Internal-Energy} + \text{Complex Energy}) \quad (3)$

The first term denotes for hydrogen bonding energy, which is determined by taking the combinations of all donor and acceptor atoms whether they form hydrogen bond. The internal energy of the ligand, which is the sum of the ligand steric and torsional energies, is calculated by using molecular mechanics expressions. The steric energy was calculated using a 6-12 potential of the form:

$$E_y = \frac{C}{d_y^{12}} - \frac{D}{d_y^6} \quad (4)$$

The torsional energy $E_{ijkl}$ is calculated by using the Tripos force field of the form:
CHAPTER 3  
INTRODUCTION

\[ E_{ij} = \frac{1}{2} V_{ij} \left[ 1 + \frac{n_{ij}}{n_{ij0}} \cos \left( \theta_{ij} \right) \right] \]  \hspace{1cm} (5)

The last term is a pair wise energy obtained for the steric energy of interaction between the protein and the ligand. A 4-8 potential of the following form with linear cut-off was used to calculate the interaction energy between the ligand and the protein.

\[ E_y = \frac{A}{d_y^8} - \frac{B}{d_y^4} \]  \hspace{1cm} (6)

The cut off distance used was 1.5 times the sum of the van der waals radii of the atoms. The 4-8 potential was parametrized to reproduce the minimum of the more usual 6-12 potential.

3.6 Applications of Docking

Over the last few years a vast amount of effort has been directed toward developing efficient docking methods and scoring functions as tools for the identification of lead compounds. Considerable progress has been made in the computational prediction of ligand-target binding modes. A number of review articles in this emerging area of research have been recently published (Taylor et al., 2002) The success application of DOCK includes the in silico virtual high throughput screen for high affinity cytochrome P450cam substrates (Taylor et al., 2002) and the computer-assisted design of selective imidazole inhibitors for cytochrome p450 enzymes (Halperin et al., 2002). Based on computational docking experiments, aminoglycosides were identified as being capable of binding the standard A-RNA duplex but not the B-DNA form. Structural evidence based on NMR solvent isotope shift measurements indicated that lividomycin, a compound suggested by the calculation, bound to the RNA major groove, corroborating the docking results. In addition, lividomycin caused a significant increase observed in the stability of the RNA duplex. Filikov (Filikov et al., 2000) identified lead compounds that disrupt HIV-1 TAR T at binding, an interaction necessary for viral replication. Very recent studies from the James group on the same target using DOCK and ICM with an improved scoring scheme produced a sub-micro molar
lead with a novel chemotype that demonstrated anti-HIV activity in a cellular assay. Using docking tools Anuradha and her co-workers have identified potent inhibitors for *Mycobacterium tuberculosis* MurC enzyme (Anuradha et al., 2010). These inhibitors are effective against resistant strains of TB.