Quality assessment of protein models

ARJUN RAY

Licentiate Thesis
Stockholm, Sweden 2011
Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av licentiate examen i biologisk fysik fredag den 9 mars 2012 klockan 10.00 i sal FB53, AlbaNova universitetscentrum, Kungl Tekniska högskolan, Stockholm.

© Arjun Ray, March 2012

Tryck: Universitetsservice US AB
Abstract

Proteins are crucial for all living organisms and they are involved in many different processes. The function of a protein is tightly coupled to its structure, yet to determine the structure experimentally is both non-trivial and expensive. Computational methods that are able to predict the structure are often the only possibility to obtain structural information for a particular protein.

Structure prediction has come a long way since its inception. More advanced algorithms, refined mathematics and statistical analysis and use of machine learning techniques have improved this field considerably. Making a large number of protein models is relatively fast. The process of identifying and separating correct from less correct models, from a large set of plausible models, is also known as model quality assessment. Critical Assessment of Techniques for Protein Structure Prediction (CASP) is an international experiment to assess the various methods for structure prediction of proteins. CASP has shown the improvements of these different methods in model quality assessment, structure prediction as well as better model building.

In the two studies done in this thesis, I have improved the model quality assessment part of this structure prediction problem for globular proteins, as well as trained the first such method dedicated towards membrane proteins. The work has resulted in a much-improved version of our previous model quality assessment program ProQ, and in addition I have also developed the first model quality assessment program specifically tailored for membrane proteins.
“There are only two ways to live your life. One is as though nothing is a miracle.
The other is as though everything is a miracle.”

*Albert Einstein*

*To my family.*
List of publications

This thesis is based on the following publications:


Paper 2. Arjun Ray, Erik Lindahl and Björn Wallner. Evolutionary information and multiple sequence alignments improve protein model quality prediction. (Manuscript)

Author’s contributions to the papers

For paper 1, B.W. suggested the project and prepared the dataset. A.R. and B.W. parameterized and optimized the method. A.R., E.L. and B.W. did the analysis. B.W., E.L., and A.R. wrote the paper.

For paper 2, B.W. suggested the project and prepared the dataset. A.R. and B.W. parameterized and optimized the method. A.R., E.L. and B.W. did the analysis. B.W., E.L., and A.R. wrote the paper.
# Contents

1 Molecules of Life ................................. 1
  1.1 Overview of proteins .......................... 1
  1.2 DNA and formation of proteins .................. 1
  1.3 Evolution and mutations ........................ 3
  1.4 Amino acids: Building blocks and their basic structures 4

2 Protein Structure ................................ 8
  2.1 Degrees of freedom in protein structure ........ 8
    2.1.1 Rotamers and side-chain rotations ........ 8
    2.1.2 Backbone rotations, torsion angles and the Ramachandran plot .... 8
  2.2 Hierarchal division of protein structure ........ 9
  2.3 Secondary structure .......................... 9
    2.3.1 α-helices .................................. 9
    2.3.2 β-sheets .................................. 10
    2.3.3 Loops ..................................... 11
  2.4 Elementary interactions in proteins ............ 11
    2.4.1 Electrostatic forces ........................ 12
    2.4.2 Van der Waals interactions ................. 12
    2.4.3 Hydrogen bonds ............................. 12
    2.4.4 Entropy and free energy .................... 13
  2.5 Anfinsen’s dogma .............................. 13
  2.6 Levinthal’s paradox ............................ 13
  2.7 Protein Folding in vivo and in vitro ............ 14
    2.7.1 Experimental Techniques .................... 14

3 Protein Structure Prediction ...................... 16
  3.1 Background and Methods ........................ 16
    3.1.1 Bioinformatics ................................ 16
    3.1.2 Alignment .................................. 16
    3.1.3 BLAST ..................................... 18
    3.1.4 Multiple Sequence Alignment ................ 18
    3.1.5 Profiles ................................... 19
    3.1.6 PSI-BLAST ................................. 19
  3.2 Structure Comparison ........................... 19
    3.2.1 RMSD ...................................... 19
    3.2.2 S-score ................................... 20
    3.2.3 GDT ....................................... 20
  3.3 Modeling ....................................... 20
    3.3.1 Homology Modeling .......................... 22
    3.3.2 Ab-initio .................................. 22
  3.4 Model Quality Assessment of Protein Models ...... 22
    3.4.1 ProQ ....................................... 23
    3.4.2 Global Quality and Local Quality of a Model ............. 23
    3.4.3 Importance of MQAP .......................... 23
3.5 CASP ................................................................................. 24

4 Machine Learning Technique 25
  4.1 Why teach a machine ? ...................................................... 25
  4.2 Neural Networks ................................................................. 25
  4.3 Support Vector Machines .................................................... 26

5 Paper Summary 28
  5.1 Model quality assessment for membrane proteins based on support vector
      machines (Paper 1) ........................................................... 28
  5.2 Support vector machine based model quality assessment for globular pro-
      teins (Paper 2) ............................................................... 28

6 Final thoughts 30

7 Acknowledgment 31
1 Molecules of Life

1.1 Overview of proteins

The cell is the most fundamental building block of all living organisms, and the definition dates back to Robert Hooke in 1665. It houses many important sub-units and molecules, in particular proteins. Proteins are responsible for many functions such as the oxygen transport by hemoglobin; they constitute antibodies for our immunity, enzymes in biochemical reactions, etc. Proteins consist of strings of elementary blocks, amino acids, folded in a three-dimensional conformation. Diseases like diabetes (Madiraju and Poitout, 2007), depression, cardiovascular defects and many others can be attributed to malfunctioning proteins.

1.2 DNA and formation of proteins

DNA or deoxyribonucleic acid is the double stranded helix of two long polymer chains of repeating units called nucleotides, containing our hereditary material. A sugar and phosphate backbone attaches one of the four bases: adenine (A), thymine (T), guanine (G) and cytosine (C). The bases form hydrogen bonds in specific pairs; adenine with thymine, and guanine with cytosine. Almost every cell in our body contains a copy of the complete human genome, which contains more than 3.1 billion of DNA bases. In nature however, larger size does not necessarily mean more complex, at least not in the case of DNA base counts. Mouse has 2.7 billion bases (Waterston et al., 2002), which is very close to human, while a simple flower such as lily can have over 100 billion bases (Bennett et al., 2004).

The formation of the protein starts with the DNA nucleotides. In the DNA, a specific combination set of three bases codes for an amino acid. This triplet is called a codon. With three nucleotides in a codon and DNA containing four types of nucleotides, the total number of codons is $4^3$ or 64 (Fig. 1).

![CODON](image)

Figure 1: A DNA codon consisting of three bases on the backbone of phosphate and sugar.

A gene is a section of the DNA containing the code for production of a specific protein. In a gene, there are encoding sections, called exons, non-coding sections called introns, and other regions such as promoters, terminators, etc. The complete set of all the hereditary material, both coding and non-coding sequences, is also called the genome. The Human
Genome Project was a global effort to sequence the complete human genome (Barnhart, 1989; Lander et al., 2001). Finding all the genes would uncover the corresponding proteins and function based similarities between them (Bentley and Parkhill, 2004).

Another molecule, which plays a leading role in the protein synthesis process, is the RNA or ribonucleic acid. Though quite similar to the DNA, it has a few key differences. The RNA consists of similar bases as in the DNA with the change of uracil (U) instead of the thymine (T). Another difference is that the RNA is a single stranded polymeric chain of nucleotides, compared to the double stranded DNA. The DNA and a few flavors of the RNA are involved in the formation of proteins. Protein synthesis can broadly be divided into two stages: DNA to RNA and RNA to protein. The schematic pathway of protein synthesis is shown in Fig 2.

The first step in this process is the expressing of the DNA chain into messenger-RNA or mRNA (Fig. 3), in a process called transcription. The genetic information of the mRNA is then translated into a chain of amino acids, which will later form a protein. This step, also called translation, is initiated by the binding of the mRNA to a translation machinery called the ribosome (Fig. 4). The next step in the translation is the binding of the first tRNA. tRNA is another specific type of RNA that binds single amino acids and contains a three bases long recognition site (anticodon) for each codon in the mRNA. There are different tRNAs for the different amino acids, each containing a unique anticodon label. A tRNA can only bind to one amino acid, but the amino acids can bind to different tRNAs containing different anticodons. The amino acid alanine can for instance bind to the tRNA containing CGA, CGG, CGU, or CGC anticodons.

The translation process is always initiated with the codon on the mRNA corresponding to methionine tRNA, i.e. AUG. After the initiation of the translation, the second tRNA anticodon binds to the next three bases on the mRNA. After the second binding, the methionine transfers from its tRNA and forms the first peptide bond with the newly arrived amino acid. Similarly, repeating the process of sliding the ribosome over
the mRNA and binding of corresponding tRNA-amino acids results in the polypeptide elongation. The chain formation will keep going on until it hits a stop codon (UAA, UAG or UGA) on the mRNA.

1.3 Evolution and mutations

The same four DNA bases are common to all the diverse life forms on this planet. This diversity can be explained from random changes in the genetic code, called mutations. Contrary to science fiction, mutation can sometimes be a beneficial process and it occurs continuously in nature. The different colors of roses or differences in body structure between humans are examples of natural variation caused by mutations.

There are different types of mutations. A substitution mutation is the exchange of a single base for another. Such point mutation is accredited with the cause of diseases like sickle cell anemia and $\beta$-thalassaemia. The insertion mutation adds bases into the DNA while the deletion would remove a whole nucleotide. Both these mutations, insertion and deletion, cause a frame-shift if it occurs in a coding region and severely alters the protein formed by the new shifted codon triplets. Diseases like Huntington’s disease are caused by these types of mutations (Davies and Rubinsztein, 2006).

One of the cornerstone concepts of biology is homology (Donoghue, 1994; De Beer, 1971). The etymology of homology lies in the Greek word homologia; meaning agreement. Homologs are units of phenotypic transformation that share common ancestry and shared developmental mechanisms (Brigandt and Griffiths, 2007). Brigandt argued that homologues are the units of evolvability (Brigandt, 2002). Hence, the study of homology is imperative in biology.

Two homologous genes can be orthologous or paralogous in nature. The orthologous
genes arise due to a speciation event, i.e., when a single species diverge into two separate ones (Fitch and Margoliash, 1970). These two genes then evolve along separate pathways, though mostly conserving the same function. An interesting example of an orthologous gene is the llama hemoglobin, which is slightly different from the human one because of the necessity to bind the relatively low concentration of oxygen in the thin air. Two paralogous genes, on the other hand, are created due to a gene duplication event, i.e., a single gene is duplicated and the event happens in a single organism. This generally results in the creation of new functionality. Myoglobin and hemoglobin are considered paralogs, where the first is an oxygen storage protein in mammals and birds (Kooymann and Ponganis, 1998) and the latter an oxygen transporter protein.

1.4 Amino acids: Building blocks and their basic structures

Proteins are chains of amino acids linked by peptide bonds. There are twenty different amino acids used in normal proteins, encoded by the DNA. The first amino acid to be discovered was asparagine (Vauquelin and Robiquet, 1806). The linear structure has a carboxylic group (C=O) and amino group (N-H) bound to Ca to form the backbone and is linked with a peptide bond to the adjacent amino acids (Fig. 6). During protein synthesis, the carboxylic group of an amino acid molecule releases a water molecule and forms a bond with the adjacent amino group. This bond is called the peptide bond. Thus, the two ends of the chain have unbound amino and carboxylic groups, respectively, which are called the N- and C-terminals of the amino acid sequence.

![Amino group and Carboxyl group](image)

Figure 6: The basic structure of an amino acid contains a carboxylic group, an amino group and a side chain.

Amino acids typically only differ in the side chain group. Table 1 lists the names of the amino acids and their one-letter-codes that will be used in the later chapters. The twenty amino acids are categorized by the interaction nature of the side-chain group (R) with water. Q,A,P,V,L,I, and M contain non-polar groups. The polar groups of amino acids can further be categorized into charged and uncharged polar moiety: S,T,C,N, and Q are uncharged polar amino acids while K,R,D, and E are charged. One more group of F,Y, and W is relatively non-polar with an aromatic side chain. Tyrosine (Y) and tryptophan (W) are relatively more polar compared to phenylalanine (F).
<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Three-letter code</th>
<th>One-letter code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>Glu</td>
<td>E</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
</tr>
</tbody>
</table>

Table 1: One and three letter code for amino acids.

There are two main classes of proteins: globular proteins and membrane proteins. Globular proteins are water-soluble proteins. The linear amino acid sequence tends to fold into a compact three-dimension structure. The name comes from the compact globin fold, identified in the first ever determined three-dimensional structure of a protein (Kendrew et al., 1958), myoglobin, which consists of eight helices. Two classic examples of proteins in this class are hemoglobin and myoglobin. Membrane proteins reside in and around the cellular lipid bilayer membrane. The cell membrane acts as a barrier between the cytoplasm, the thick solution in the interior of the cell containing all the organelles, and the outside. Membrane proteins include signal transduction proteins, ion flow proteins, anchor proteins for other molecules, etc. Membrane proteins have at least one trans-membrane segment spanning the bilayer. They contain both hydrophobic as well as hydrophilic amino acids on their surface, making them very difficult to work with in all experimental aspects. An example of membrane proteins is the class of ion channel proteins that, as the name suggests, are involved in transport of ions. Ion channels are responsible for exchange of ions between the extracellular and intracellular regions.

Life began billions of years ago with the origin of the prokaryotes, single cell organisms that lack a nucleus (Zimmer, 2009). This means that the DNA and other materials existed throughout the cell. Eventually the cells developed a membrane, a universal selective barrier. The role of the membrane was initially to adsorb proteins and DNA/RNA from the surrounding environment. In more recent times, in addition to the role of simple compartmentalization, membranes adsorb nutrients in the intestine, they encapsulate nerves as in myelin, they act as anchors for the cytoskeleton for shape, they capture light.
in the rod cell membrane, and maintain charge and concentration gradients within the cell (Luckey, 2008).

The cellular membranes consist of lipids, carbohydrates and proteins. Lipids are molecules having a hydrophilic head-group and a hydrophobic tail end. The membrane is made up of a bilayer of phospholipids. The polar groups face the extracellular and cytosolic part while the hydrophobic acyl groups face inwards. The aggregation of lipids into layers can be explained through the hydrophobic effect. Water disfavors non-polar molecules and tends to form a cage around the non-polar moiety. This also leads to restricted mobility for the water molecules. Aggregation of the non-polar groups thus leads to less surface area for the non-polar acyl groups to interact, resulting in a decrease of the number of immobile water molecules forming the cage. The separation in the bilayer is thus hydrophobically favorable where the acyl-water interaction surface is minimized (Tanford, 1980).

Proteins interact with the lipid bilayer in many ways. The peripheral proteins attach on the surface of this bilayer while inserted proteins are present in the non-polar region of the membrane. The lipids as well as the proteins together play a critical role in permeability of ions and other polar/non-polar substances. The percentage of protein versus lipids varies a lot in a membrane, and it is believed that different protein/lipid ratios or lipid compositions can play important roles for membrane protein function (Cullis and De Kruijff, 1979).

![Fluid mosaic model of the cell membrane](image)

Figure 7: Fluid mosaic model of the cell membrane. Adaptation from Mind the membrane (Pietzsch, 2004).
In 1972, Singer and Nicolson established a model, called the fluid mosaic model of the cell membrane (Singer and Nicolson, 1972), to illustrate the cellular membrane and its features. The “mosaic” alludes to the proteins scattered in and on the lipid bilayer. The membrane is not a static barrier, simply consisting of lipids, but a highly dynamic body containing a mixture of proteins, lipids and carbohydrates (Fig. 7).
2 Protein Structure

2.1 Degrees of freedom in protein structure

The polypeptide chain of the amino acids is essentially a long chain of carbon and nitrogen in the backbone and side chains branching out for each amino acid. As the carbon-carbon bonds in this chain are rotatable, the polypeptide chain is quite flexible, allowing it to adopt almost any configuration. There are also additional degrees of freedom in the bond length and bond angles, which can vary slightly around their average values. The rotations around the carbon bonds in the backbone have the largest impact on the structure, while the rotations of the carbon bonds in the side chains only affect that specific side chain.

2.1.1 Rotamers and side-chain rotations

Conformational isomers are compounds having the same chemical formula but difference in structure due to rotation along a single bond, corresponding to distinct potential energy minima. These isomers are also called rotamers (Moss, 1996). In a small protein, a deviation of more than 20° of such rotamers can yield energy differences $\sim 2$ kcal/mol as well as geometric differences of 0.5 Å (Schrauber et al., 1993). The side chains would prefer to maximize the distance between the groups on the two adjacent carbon atoms so that they adapt the most energetically favorable conformation. This usually results in a staggered conformation. For each carbon-carbon side chain bond, there are three such possible conformations separated by angles of 120°. The discretization of states for each degree of freedom is beneficial for predicting and developing prediction methods for proteins.

Rotamer libraries are collections of rotamer information about the side chains’ dihedral angles at a residue level and are widely used in protein structure prediction, protein design, and structure refinement. With the increase of structural data in recent years, it has become possible to derive well-refined rotamer libraries for data inclusion and for studying the dependence of the rotamer populations and the dihedral angles on local structural features (Dunbrack, 2002). A popular method, SCWRL, predicts the conformations of protein side-chains, starting from main-chain coordinates alone, using a rotamer library (Bower et al., 1997).

2.1.2 Backbone rotations, torsion angles and the Ramachandran plot

The polypeptide chain of amino acids is covalently bonded in a $-C_\alpha-C-N-C_\alpha-$ manner. If the next carbon atom after N and $C_\alpha$ is denoted as $C'$, there are three possible rotations around the bonds in the backbone, around $-N-C_\alpha-$, $-C_\alpha-C'$- and $-N-C'$-. These three angles of rotations are called $\phi$ (phi), $\psi$ (psi) and $\omega$ (omega) respectively. The variation of the $\phi$ and $\psi$ results in formation of secondary structures. Rotameric position of an amino acid’s side chain can also affect the backbone $\phi$ and $\psi$ angles (Dunbrack and Karplus, 1994).

The flexibility in the $\phi/\psi$ angles leads to significant conformational freedom in the unfolded state. It is reduced to a single state, having no steric clashes, in the native structure. A few features might reduce the conformational flexibility. Two cysteine side
chains in the proximity of 2 Å can form a covalent bond between the two sulphurs. The disulphide bridge plays an important role in protein folding by limiting the conformational freedom of the unfolded protein molecule, especially of short polypeptide chains. Intermolecular disulphide bonds between two individual peptide chains or proteins, such as in insulin, result in similar stability. Another example is proline whose side chain contains an extra covalent bond with the backbone of the polypeptide chain, which results in a fixed dihedral angle. This rigidity means less conformational freedom in the unfolded state. Proline is critical for thermo-stability of proteins in thermophiles (Suzuki et al., 1987; Friedman and Weinstein, 1965). Thermophiles are organisms living in extremely hot conditions such as inside a volcano, where the conditions put extreme requirements on protein structure.

Rotations about the different torsion angles are intrinsically similar and governed by energetics. The trans-form of a peptide bond ($\omega = 180^\circ$) is favored by a ratio of 1000:1 over cis-form ($\omega = 0^\circ$). This is because the steric clashes in the cis-form makes it energetically unfavorable (Mathieu et al., 2008). Existence of both isomers is difficult unless the inter-conversion is rapid. Proline is however an exception, where the cis-form is favored due to the extra bond between the backbone nitrogen and the side chain (Stein, 1993). Prolyl peptidyl bonds have shown to reverse from cis (7%) and trans (93%) in the unfolded state to cis (85%) in the N-terminal of G3P of the filamentous phage fd (Jakob and Schmid, 2008).

The Ramachandran plot is a way to visualize the sterically allowed $\psi$ and $\phi$ angles of amino acids in a protein structure (Ramachandran et al., 1963). The fraction of conformations that are fully favorable (apart from proline and glycine) is merely $\sim 7.5\%$ (darker blue) while the partially allowed region (lighter blue) is $\sim 22\%$ (Fig. 8).

### 2.2 Hierarchal division of protein structure

Protein structure can be divided hierarchically into four structure types: primary, secondary, tertiary and quaternary structure (Linderstrom-Lang, 1952). The primary structure is the linear sequence of amino acids. The secondary structure is a local three-dimensional motif of a section of the protein. The tertiary structure describes the arrangement and interactions between secondary structures. Finally, the quaternary structure is the configuration of one or several tertiary structures that together combine to form a functional complex.

### 2.3 Secondary structure

Secondary structure is defined by the torsion angles of the backbone. $\alpha$ helices and $\beta$ sheets are two classes based on the local backbone geometry. The populated regions in the Ramachandran plot can be used to illustrate this. The left hand upmost region and the middle left region are indicative of secondary structures in proteins.

#### 2.3.1 $\alpha$-helices

One of the simplest ways to locally pack a polypeptide chain is to curl it up as a helix, which also makes it possible for the amino acids to form stabilizing hydrogen bonds. This
common structure is called α-helix, and was first suggested by Linus Pauling (Pauling and Corey, 1951). The N-H group of an amino acid forms a hydrogen bond with the C=O group of the amino acid four residues before it, in an $i \rightarrow i + 4$ fashion. Due to the geometry of the helix, the hydrogen bonds are all oriented along the helical axis. This gives the helices a dipole moment because of the arrangement of the ionic bonds. The hydrogen bonding with different neighboring residues govern the classification of the α-helices. In nature, only right-handed helices are found. The α-helix has 3.6 residues per turn.

2.3.2 β-sheets

β-strands are pleated in shape. The side chains of the adjacent residues are in the opposite direction and the backbone has an angle of 120°, giving rise to the curvy form. The hydrogen bonds are formed between the C' = O and the NH groups of two adjacent β-strands. Extending this pattern of hydrogen bonding between the strands results in a pleated sheet structure. This is called a β-sheet. The β-sheet formation involves long-range interactions between residues that are close in terms of space yet are distant in the primary sequence. A β-sheet can be described roughly by giving the number of strands, their topology, and their direction of the backbone, running parallel or antiparallel. For parallel sheets, the hydrogen bonds are evenly spaced out in a certain angle orientation while for antiparallel sheets, they alter between narrowly and broadly separated pairs (Branden and Tooze, 1999).
2.3.3 Loops

The parts of the polypeptide connecting the well-defined alpha helices and beta sheets are called loops and are typically less structured and flexible. The loops usually occur on the surface of the protein and frequently contain polar residues that interact with water. Mutations in the loop regions are often tolerated since they do not break the core of the protein. For instance, this might be used to alter the binding specificity. Hairpin loops are the units that connect two adjacent antiparallel $\beta$-sheets and are sometimes simply called turns. In small loops, a single cis-proline or glycine has been found to stabilize and promote the folding process. With the lack of a $C_\beta$ atom, glycine has a faster rate of contact formation than other amino acids due to the increased backbone flexibility. Cis-proline contributes to a faster loop formation because of its largely restricted conformational space arising from its cyclic side chain (Krieger et al., 2005).

Fig. 9 shows the various secondary structures discussed above. The coil is the $\alpha$-helix and the arrow-heads represent a parallel $\beta$-sheet. These two are connected by the loops.

![Structure of a mutant KcsA potassium channel with its secondary structures. PDBID: 1ZWI.](image)

Figure 9: Structure of a mutant KcsA potassium channel with its secondary structures. PDBID: 1ZWI.

2.4 Elementary interactions in proteins

The sequence of a protein is believed to be solely responsible for the resulting native structure. Covalent bonds are formed by the interactions of electron pairs and are the strongest chemical bond in a protein. The cysteine amino acids forming disulphide bonds, peptide bonds, and all bonds between atoms in the amino acids itself are all covalent interactions. Several other elementary forces also play a major role in the folding process. Enthalpy contributes via the total interactions within the polypeptide chain.
2.4.1 Electrostatic forces

Non-covalent interactions are mostly interactions between charged ions. The electric potential between two point charges, $q_1$ and $q_2$, at a distance $r$, is described by Coulomb's law

$$E = \frac{q_1 q_2}{4\pi\epsilon r}$$  \hspace{1cm} (1)

The Coulomb interactions are long-ranged as they decay as $1/r$. The interaction of $q_1$ and $q_2$ is repulsive for charges of the same sign, and attractive for opposite-sign ones. $\epsilon$ is the absolute permittivity. Ionic bonds, between oppositely charged groups of amino acids in close contact, are extremely favorable.

2.4.2 Van der Waals interactions

All atoms or molecules attract each other. Neutral atoms or molecules are made up of charged particles. The net cancellation of these charges present a neutral picture. The attractive force generated by the shift of electrons towards another nucleus, results in polarization of the charge. As two charges are brought nearer and nearer, the short-range repulsive force grows exponentially. The intermolecular attractive force is known as Van der Waals interaction.

The long-range attraction decays as $1/r^6$. We can model the short-range repulsion by $1/r^{12}$. $1/r^{12}$ can be obtained in a single multiplication once $1/r^6$ is available and is hence cheap to calculate.

$$U_{ij,LJ} = 4\epsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right)$$ \hspace{1cm} (2)

where $U_{ij,LJ}$ is the Lennard-Jones potential for a pair of sites $i$ and $j$ separated by a distance $r_{ij}$. $\sigma_{ij}$ and $\epsilon_{ij}$ are Lennard-Jones parameters.

2.4.3 Hydrogen bonds

The pull of the electron towards a more electronegative atom along a bond will polarize the pair of atoms, and create a larger partial charge on the hydrogen. When this hydrogen is close to a different electronegative atom, we will get a hydrogen bond between two different molecules due to the strong polarization of both. The energy of such a bond is around five kcal/mol and has short-range interaction. Charged side chain groups are mostly present on the surface of proteins and facing the aqueous phase. This leads to hydrogen bonding with the medium, which also reduces the effective charge of such groups.

The C=O and N-H groups in polypeptide chains easily form hydrogen bonds. In the case of protein molecules in aqueous solution, amino acid residues would release two water molecules, previously bound to them and form a hydrogen bond. Hydrogen bonds are common in the secondary structure of protein.
2.4.4 Entropy and free energy

Although all individual conformations are equally probable, there are lots of conformations that are considered unfolded or unordered, but very few ordered conformations. Even if they were to have exactly the same enthalpy, this probability would factor in and make it more likely to find the system unordered than ordered. Moving unordered to ordered conformations is more unfavorable the higher the temperature is. It is possible to show that the most favorable conformation is the one that minimizes the expression of the free energy

\[ F = E - TS, \]

where \( F \) is the free energy, \( E \) is the energy and \( S \) is the entropy of the state at a given temperature \( T \).

The folding and unfolding reactions of a protein are governed by the entropy and the non-covalent interactions (Zondlo, 2010). The hydrogen bonds have energies that are in the range of 2-5 kcal/mol and Van der Waals interactions energies are approximately in the range of 0.01 to 0.2 kcal/mol. The seemingly small values of energies adds up to direct the fate of the folding pathway. Small changes in pH or temperature can result in destruction of the secondary and tertiary structures of a folded structure of protein and is indicative of these weak non-covalent interactions.

2.5 Anfinsen’s dogma

Christian B. Anfinsen proposed that the native structure of a protein is determined solely on the basis of its constituting amino acid sequence, for which he was awarded the Nobel Prize in chemistry in 1972. He showed this by first fully denaturing an enzymatic protein, ribonuclease, containing 8 S-H groups. The mixture of products formed was inactive and retained the total count of disulphide bonds. When this process was reversed, using the products in the mixture, it converted into a homogeneous product, indistinguishable from native ribonuclease (Anfinsen, 1973). This led Anfinsen to postulate that at a specific condition, such as temperature and pH, the protein adopts the structure with the lowest free energy. This is known as Anfinsen’s dogma (also called the thermodynamic hypothesis).

2.6 Levinthal’s paradox

The polypeptide chain can fold into a protein, having a number of different conformations. The total number of possible conformations is of astronomical magnitude. Let us consider a simple example - a protein of 100 amino acids having 99 peptide bonds. This will result in 198 different degrees of freedom for its \( \phi \) and \( \psi \) angles. Assuming three possible states for each angle, the total number of possible conformations is \( 3^{198} \). Even if nature could sample each of these conformations in a matter of femtoseconds, the total time to exhaust all the possibilities would be more than the age of the universe!

In 1969, Cyrus Levinthal famously alluded to this by recognizing the fact that though proteins fold spontaneously and on a short time scale (in the order of milli/micro-seconds), the astronomical number of possible conformations of a polypeptide chain makes it a
paradox that folding into a protein still occurs (Levinthal, 1969). He suggested that the paradox can be resolved if the rapid formation of local interactions can guide and speed up the protein folding. The local amino acid sequences which form stable interactions can serve as nucleation points in the folding process (Levinthal, 1969). Levinthal suggested a well-defined kinetically controlled pathway to the native structure (Levinthal, 1968; Dill and Chan, 1997).

2.7 Protein Folding in vivo and in vitro

Protein folding is considered to be governed by long range interactions within a protein (Ngo et al., 1994). Short-range interactions would not have resulted in such a hard and complex problem (Fraenkel, 1993; Unger and Moult, 1993). The free energy difference between the unfolded, denatured state and the native structure is not more than the energy of a hydrogen bond. This makes the folding process a precarious balance. The heterogeneous contact combination of different residues in a polypeptide chain decides the final outcome of the folding process (Plotkin and Onuchic, 2000). The end result of the folding process of the polymer chain is the lowest possible free energy state. The stability of the native folded protein is affected by the mutations of the amino acids (Zeldovich and Shakhnovich, 2008; Onuchic et al., 1997). Mutations happen on a regular basis. Natural selection disfavors mutations that negatively affect a protein’s function (Onuchic et al., 1997).

The polypeptide chains of amino acids that are longer than 100 residues have a tendency to fall back among themselves and form non-native aggregates (Brockwell and Radford, 2007). These are called folding intermediates. These intermediates can either be kinetically stable enough that the non-native state needs to be reorganized through some mechanism or can be stepping stones on the path to reach the final native state. Biomolecules, classified as chaperonins, assist the process of folding the intermediate step into the native state by binding to the unfolded, partially folded and misfolded proteins (Evstigneeva et al., 2001).

Other kinetic factors such as incorrect disulphide bonds have the property to stabilize a structure due to the reduction in conformational entropy, i.e., the entropy arising from different conformations of the backbone. These kinetic factors can contribute against the folding process towards the native structure. Protein folding is driven by entropy as well as enthalpy. As mentioned above, the final resultant of the folding process is the lowest free energy state. The energy landscape of this funnel is rough, since it contains multiple non-native local minima. There are multiple minima separated by large barriers, defining the meta-stable states (Fig. 10).

2.7.1 Experimental Techniques

Experimentally determining structures can be a time consuming process. The two most common techniques used are X-ray crystallography and nuclear magnetic resonance (NMR).

X-ray crystallography is a method in which beams of X-rays collide on a crystal and diffract in specific directions, based on the arrangement of atoms within the crystal. A pure protein in an aqueous solution precipitates and crystallizes under specific conditions such as pH, concentration of protein and temperature. A well-defined crystal of a
protein will yield a high resolution picture of the spatial arrangement of the individual atoms (Chernov, 2003; Rupp and Wang, 2004). The limitation of this method lies in the difficulty of forming the crystals. In addition, the picture obtained from this technique is just a single snapshot of a certain protein. Slightly different conditions can lead to a difference of as much as 1 Å of the native structure between two snapshots of the same molecule of protein.

NMR is based upon the magnetic spin behavior of the atomic nuclei of a molecule. Each nucleus of atoms experiences a unique chemical shift because of the specific chemical environment when introduced in a magnetic field. The structure can be calculated by introducing the molecule in a high-intensity magnetic field and studying these chemical shifts. The biggest advantage this method has over X-ray crystallography is its ability to study the protein in the solvated state rather than as a crystal. The trade-off is that NMR results in lower resolution of the protein. NMR studies also require a high concentration of protein.

Figure 10: The energy landscape of protein folding.
3 Protein Structure Prediction

3.1 Background and Methods

3.1.1 Bioinformatics

A term initially coined by Paulien Hogeweg and Ben Hesper in 1978 (Hogeweg, 1978); bioinformatics describes the comparative analysis of genome data by computational methods (Hogeweg, 2011). This discipline has three key avenues of research: organizing biological data in databases, analyze this data, and, perhaps the most important, extracting biological importance out of the data (Luscombe et al., 2001).

There has been an explosion of biological data in the last couple of years (Choi et al., 2007; Reichhardt, 1999). As of April 2011, the NIH genetic sequence database GenBank (Benson et al., 2008) has ~ 135 million sequences mapped with more than 12.6 billion bases, doubling every 18 month. Even if the present-day databases contain a significant amount of data, this is only the beginning. The pace of sequencing is increasing at a rate even modern computer technology has trouble to keep up with. The reason is that the cost of sequencing has gone down considerably (Fig. 11). The first human genome cost ~ 3 billion USD. Now it is possible to sequence an individual human for a fraction of that initial cost. This represents a fundamental challenge in bioinformatics; both from an analysis and a method development point of view.

![Baseline information](image)

Figure 11: In the last decade, the cost of DNA sequencing has fallen to a hundred-thousandth of its initial cost (Carr, 2010).

3.1.2 Alignment

A challenging question in molecular biology is to what extent a protein’s function can be determined from its sequence, since proteins with high sequence identity frequently
have similar functions. It has been shown that above a threshold of around 50% residue identity, the functional divergence is decreased dramatically (Sangar et al., 2007). Comparing sequences becomes critical when doing evolutionary studies as well as functional or structure predictions of proteins. In this comparison, it is common to find proteins of the same family having positions or even a block of mutated residues. The goal of sequence alignment is to identify residues in two sequences that correspond to each other (Lesk, 2002). Alignment algorithms measure the similarities by observing the patterns and variations of conservation and draw conclusions about their evolutionary relationships and mutations between sequences (Mount, 2004). The sequences with high identity have obvious functional similarities while the functional similarities for too low identity sequences are very difficult to detect. When two sequences have more than 30% residue similarity, they are typically homologous. This range is known as the twilight zone (Rost, 1999).

We need to quantify the degree of biological relationships between sequences, amino acids or nucleotides. The mutated amino acids can be substituted by amino acids of similar or completely different chemical nature. The former is frequent, but the latter much rarer since it can affect the protein’s structure and function. To properly align two sequences, we need a scoring method that takes this into account. It should be used to account for the chemical nature of amino acids. The scoring matrix is on the amino acid/nucleotide level. It gives values to represent the residue similarities. Two important scoring matrices are PAM (Dayhoff et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992). Point Accepted Matrix or PAM was one of the first amino acid substitution matrix scoring methods. It was built by global alignment of sequences from 71 groups of 34 superfamilies, with similarity above 85%. For each group, a phylogenetic tree is constructed where the evolutionary distance of one PAM is equal to the probability of one point mutation per hundred residues. It is based on a Markov model where each change depends on the current state, but not the history of previous states. The drawbacks of PAM, in terms of assuming equal mutation rate for all amino acids in a sequence and an outdated dataset, makes BLOSUM a more attractive candidate for a scoring scheme. BLOcks SUBstitution Matrix or BLOSUM was developed using locally aligned proteins with 62% and above identities. The probabilities of substitutions are calculated on blocks of conserved sequences.

Two historically important algorithms developed for alignment of sequences are the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) and the Smith-Waterman algorithm (Smith and Waterman, 1981). The Needleman-Wunsch algorithm compares and tries to maximize the score based on residue-matching covering the entire length of two sequences. Comparison of all possible pairs of residues is called global alignment. This is mostly useful when the sequences are of similar lengths. The Smith-Waterman algorithm compares local segments of the sequences, instead of extending the alignment over the entire sequence. This method of comparison is called local alignment and is useful in the case of varying length of the sequence having similar regions in them. The only essential difference is that the Needleman-Wunsch algorithm reports alignment corresponding the maximum score over the whole sequence length and Smith-Waterman algorithm reports the alignment corresponding to maximum subsequence.
3.1.3 BLAST

Basic Local Alignment Search Tool or BLAST is a heuristic approach to the database search problem (Altschul et al., 1990). It has revolutionized the identification of similar sequences in large databases and has been quoted as the paper with most citations in 1990s (Taub, 2000). Instead of comparing the query sequence with all sequence in the database, BLAST splits the query sequence into a shorter word list and use this list to scan the database for hits. The database search results in a list of similar sequences, i.e. sequences with short word matches above a threshold. To get the individual residue similarity these hit are then realigned using the standard Smith-Waterman algorithm. The trick with BLAST is the word list. First, it is possible to pre-calculate a hash table for all word matches in the database to speed up the search. Second, the word list makes it possible to filter out many sequences with no similarity (no short word match) to the query without the need for a full-fledged sequence alignment.

3.1.4 Multiple Sequence Alignment

The pair-wise sequence alignment often misses subtle similarities. It is important to highlight the position conservation of residues as well as complex evolutionary divergent relationships (Lobo, 2008). In the Multiple Sequence Alignment (MSA), the homologous sequences' residues in a given column may have a common functional role (Edgar and Batzoglou, 2006). MSAs are able to identify a much more detailed and biologically relevant picture of a protein family than single sequences. The representation of frequencies of amino acids in a protein sequence of a family highlights the conserved residues (Gribskov and Veretnik, 1996). Conserved amino acid stretches have been widely used in protein structure prediction (Rost and Sander, 1994; Jones, 1999) and isolation of relevant regions facilitates a number of biological analyses (Attwood, 2002) (Fig. 12).
3.1.5 Profiles

A profile is a representation of the multiple sequence alignment of all conserved regions. It creates a substitution matrix in which each position in a group of aligned sequence is assigned a score for each of the twenty possible amino acids.

Profile-sequence analysis is a method in which one queries a sequence against a profile of conserved regions in a multiple sequence alignment. This is more effective as profiles inherently have more information than a single sequence to compare with. Yet, replacing both the target and template sequences with profiles has shown even more promises for distantly related proteins that fall under the twilight zone (Rychlewski et al., 2000; Yona and Levitt, 2002; Rychlewski et al., 2000).

3.1.6 PSI-BLAST

One powerful variant of profile-based methods builds frequency profiles using BLAST in its search and is called Position-Specific Iterated BLAST or PSI-BLAST (Altschul et al., 1997). The method of doing an initial search, constructing a profile from the hits of the search and then iterating this search based on the profile improves the sensitivity. PSI-BLAST constructs a multiple-sequence alignment with reference to a database by running BLAST, building a profile in its first step. This profile is used in place of the original substitution matrix for a further search of the database to detect sequences using profile-sequence search. The newly detected sequences from this second round of the search, which are above the specified threshold, are again added to the alignment and the profile is updated for another round of searching. A strength of PSI-BLAST is running as a start-to-finish program where one inputs the protein sequence and searches a database, with the resulting profile as a by-product after all the intermediate steps of alignment, scoring and iteration.

3.2 Structure Comparison

With the growing number of structures being determined, it is important to have methods for comparing apparently similar three-dimensional structures. Homologous proteins sharing similar three-dimensional fold have evolved beyond sequence recognition. Structure comparison is thus highly desirable and has therefore received considerable attention for many years (Shih and Hwang, 2003; Gibrat et al., 1996; Holm and Sander, 1996). A wide range of different quality measures is used in the literature to compare models structurally with the target. The aim is to quantify the correctness of the models compared to the actual native structures. Some of the methods are given below.

3.2.1 RMSD

Root mean square deviation (RMSD) measures the similarity of the atomic coordinates after optimal superposition. It is mainly sensitive in identifying poorly modeled regions of loops in a structure (Maioorov and Abagyan, 1997).

\[
RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta^2_i},
\]
where $\delta_i$ is the distance between two positions for atom $i$ and $N$ is the number of pairs of equivalent atoms.

3.2.2 S-score

If the structural model is close the native (say $<3$ Å), RMSD is a pretty good measure. However, when the structural model deviates further from the native structure the RMSD measure gives far too much importance to the residues that are bad (or merely less well-defined) compared to the ones that are actually correct. Several measures have been developed to account for this by maximizing the sum of a normalized distance, for instance

$$S_i = 1/(1 + (d_i/d_0)^2),$$

where $d_i$ is the deviation for residue $i$ and $d_0$ a distance threshold. The S-score ranges from 0 (bad model) to 1 (perfect).

3.2.3 GDT

Another measure of protein similarity that is commonly used is the GDT score (Global Distance Test). The GDT score is calculated as the largest number of $\alpha$-carbons that can be optimally superimposed, within a defined distance cutoff. To improve the sensitivity the score is typically calculated over several different cutoffs. The most used cutoffs are shown below:

$$GDT_{TS} = 100 \times \left( \frac{\sum_{d_i} (GDT_{di})}{NT} \right) / 4$$

where $d_i \in (1.0, 2.0, 4.0, 8.0)$ Å distance cutoffs.

As S-score, GDT score is more accurate compared to RMSD, in particular if the two structures that are compared contain regions that differ.

3.3 Modeling

For proteins, structures are more conserved than sequences. The function of a protein is governed by its structure. An example of how the structures become more similar with increasing sequence similarity is shown in (Fig. 13). There is an almost negligible structural difference between sequences in the range of $40 - 100\%$ sequence identity, followed by a sharp rise below $40\%$ sequence identity, indicating the start of the twilight zone.

Development of software and hardware has given us insight into protein structures at a fraction of the cost and time compared to the experimental techniques. In case of unknown structures, several methods are used to estimate the three-dimensional structure of a protein. Homology modeling is an example of such prediction of an unknown structure based on its sequence similarity to other structures. Fig. 14 shows the classic workflow of homology modeling, which is discussed below.
Figure 13: A: Models with less than 30% sequence identity to the target frequently have substantial alignment errors and a suboptimal template selection. B: Models with 30-50% sequence identity to the target frequently have alignment errors in the non-conserved segments of the target protein, structural variation in templates, and incorrect reconstruction of the loops. C: Models having higher than 50% sequence identity to the target typically have the correct fold and the alignments tends to be mainly correct.

Figure 14: Flowchart of homology modeling. Courtesy of Dr. Björn Wallner.
3.3.1 Homology Modeling

Homology or comparative protein structure modeling constructs a three-dimensional model of a given protein sequence based on its similarity to one or more known structures (Jacobson and Sali, 2004). The minor changes in the sequence result in a small change in the structure. The most essential step in homology modeling is the alignment of two or more sequences to determine whether they are homologous. It gives a very good idea about the conserved regions and can identify the functional residues. One of the biggest factors to the success of homology modeling is the knowledge of almost all protein folds. Today, 90% of the structures submitted to the PDB correspond to already known folds (Jinbo et al., 2004). This means that it is rare to find a completely new protein which has no homologous sequence or folds known.

In reality though, it is not trivial to specify strict algorithms to identify homologous proteins at low levels of sequence identity. Hence, a database search to find a homology match of a protein does not always result in giving the best pair. Using information from multiple templates can help improve the quality of homology models (Larsson et al., 2011).

3.3.2 Ab-initio

In the scenario where the target has no similarity to an already known protein structure, the process resorted to is called ab-initio structure prediction. The aim of this procedure is to predict the structure of the protein and to apply physical principles to search through the various possible solutions. Ab-initio structure prediction requires a large computational resource even for tiny proteins and methods like Folding@Home and Rosetta@Home utilize distributed computing power to solve this problem.

The basic principle governing this method of structure prediction is that the native configuration of a protein molecule is the global free energy minima for the structure. However, as previously discussed, it is an exceptionally complicated search problem, as the search space is gigantic to sample and to check all the possibilities is enormously computationally expensive.

ROSETTA (Simons et al., 1999; Bonneau and Baker, 2001) assembles smaller fragments of known proteins and tries to assemble them using a Monte Carlo sampling approach along with knowledge-based energy function. The reason for employing fragments in structure prediction is to reduce the complexity of the problem by reusing parts of known structures, which improves both performance and accuracy significantly. Successful ROSETTA predictions typically yield 3–6 Å Cα RMSD models for 60-residue proteins (Rohl et al., 2004).

3.4 Model Quality Assessment of Protein Models

It is now easy to generate a large number of models in a short period of time. Therefore, it becomes critical to be able to judge and rank these models based on their quality. The first methods in the field, like ProsaII (Sippl, 1993), Verify3D (Luethy et al., 1992) and Errat, were developed as techniques to detect erroneous crystal structures. The development of the R-free method (Brunger et al., 1987) has made these methods less important for crystallography. There has since been a paradigm shift into now using
these methods for assessing the quality of protein models instead. This has also led to the development of methods, Model Quality Assessment Programs (MQAPs), dedicated to the prediction of protein model correctness.

Methods that only use information from a single model to assess the quality of a model are called stand-alone methods. On the other hand, another genre of MQAPs gather predictions about a model’s quality from a range of different methods and then compare the results. The basic idea behind such consensus analysis is that predictions derived from a range of other methods should contain more information to compare and that a better model will have many neighboring similar models and hence yield a higher score. MQAPs using such consensus analysis are intuitively called consensus methods. These consensus MQAPs do not try to evaluate the quality of a model. Instead, they use the similarity between a model and many other models. This type of evaluation is only useful when there is a large set of models for the same target, which is typically the case in homology modeling.

### 3.4.1 ProQ

ProQ was the first method that used protein models instead of native structures, with different similarity as a target function (Wallner and Elofsson, 2003). It is an example of a stand-alone method. ProQ uses neural networks to score models from a few basic structural features such as atom contacts, exposed surface area, and predicted secondary structure. This was subsequently combined with Pcons to improve the prediction even further (Baker and Sali, 2001; Zhang, 2009). Pcons (Lundstöm et al., 2001) was the first consensus method that predicted the quality of all models, by assigning a score to each model reflecting the average similarity to the entire ensemble of models.

### 3.4.2 Global Quality and Local Quality of a Model

The scoring functions can be assigned on two levels. We can assign a single score while assessing the model as a whole (global quality) or at residue level (local quality). Global quality is beneficial in model selection processes since a single measure for each model makes it easy to qualitatively rank all models in a dataset. The importance of the local residue level quality measure becomes more prominent when there are good regions in an average/low global quality model. This gives better insight into which part of the model can be trusted for making further improvements. Fig. 15 highlights the importance of the local quality measure in a model.

### 3.4.3 Importance of MQAP

After the rise of high throughput sequencing in 2004, it is now normal to produce millions of sequences per day. A bottleneck thus becomes the determination of the three-dimensional structure from these sequences. Computational power has been rising rapidly each year and currently available methods can produce tens of thousands of model from a single sequence in a day. It then becomes critical to have the ability to pick the best model of these, which in an ideal world would be the native structure. Researchers around the globe are trying to build faster and more accurate MQAPs to meet this growing demand.
Figure 15: Predicted as well as true deviation for an early model of the $\beta_2$-adrenergic receptor. This GPCR model by Skolnick (Zhang et al., 2006) was published before any experimental structure was available, but ProQM correctly identifies regions where the model was later found to deviate from the X-ray structure (Ray et al., 2010).

The challenge now is to build good models and choosing the correct one, quickly and accurately, to complement the pace of high throughput sequencing.

3.5 CASP

Every second year since 1994, protein modelers test their methods in an international experiment called Critical Assessment of Techniques for Protein Structure Prediction (CASP). Predictors in applied mathematics, computer science, biology, physics and chemistry in well over 100 scientific centers around the world work for approximately three months to generate structure models for several tens of protein sequences selected by the experiment organizers (Kryshtafovych and Fidelis, 2009). The proteins suggested for prediction (in the CASP language: targets) are either structures on the verge to be solved or structures already solved and deposited in the PDB and withheld until the end of the CASP season. The categories in CASP 9 in 2010 involved tertiary structure predictions for template based modeling and free modeling for targets having no suitable templates to use, refinement of models by performing finer corrections to get higher quality models, identifying disordered regions detecting residue-residue contacts, function prediction and model quality assessment. CASP assesses the submitted models with GDT_TS score along with a more sensitive scoring scheme, GDT-HA, which has smaller cutoffs for the differences.
4 Machine Learning Technique

4.1 Why teach a machine?

Structure prediction of proteins can broadly be based on two methods. Physics-based methods assess a model based on the laws of physics and knowledge-based methods judge the information of a model based on the rules of evolution. One of the most important lessons from the previous rounds of CASP has been the large success of knowledge-based prediction methods. Some of the more successful knowledge-based software involved methods of learning or extracting knowledge from the existing protein structures and to implement this gained knowledge into predicting other unresolved structures (Larranaga et al., 2006). Machine learning techniques deal with pattern recognition, classification, and prediction, based on derived information from existing data (Tarca et al., 2007). It is inherently an approximation of all the knowledge fed into the training and hence has a limited accuracy.

Machine learning can successfully be used to classify models according to their quality. The core idea of machine learning is to train a program by showing it sets of good and bad examples, and subsequently use the trained program to predict other samples. Ideally, with an unknown model, the method should be able to classify it in either of the two categories by statistical fitting of the unknown example to the examples the method was trained on. The biggest strength of machine learning technique is that it is fast and that no explicit relationships are needed beforehand.

There are two alternatives for the classification approach: supervised and unsupervised classification. In supervised classification, the dataset is divided into classes, which are governed by the data’s characteristic or feature. The feature is a quantity that describes an instance and is specific to a dataset (Kohavi and Provost, 1998). An example of features would be categorizing a dataset of human faces based on the features, such as color, position of eyes and nose, or a dataset of different animals, differentiated by the various features such as habitat and size. The aim of the supervised classification problem is to accurately assign new members to different classes according to its features (Tarca et al., 2007). This labeling of classes, before the classification of a data, indicates the supervised learning. Contrary to this, in the unsupervised classification problem, a new data discovers its own grouping without any pre-assigned class structure. The algorithm would cluster similar featured data into classes based on certain similarities of attributes. Coming back to the previous example, a supervised classification would be when the animals of the animal dataset are assigned to a specific pre-assigned features, as the ones mentioned above. This is different in unsupervised classification, as the algorithm is let to study and classify the animals in the dataset. On inspection, we may find that one cluster of data is now based on the common sizes, while another one may be based on the feature of color. Hence, the data found its own grouping.

4.2 Neural Networks

An artificial neural network is the attempt to mathematically model an algorithm that mimics the human behavior of the brain to adapt and extract information (McCulloch and Pitts, 1943). A neural network will have a component for recognition of a data
and another for classification, which are arranged in layers such that the output from many connections, or nodes, is fed into a single node as input (Hirst and Sternberg, 1992; Presnell and Cohen, 1993; Rost, 1996; Rost and Sander, 1994; Larranaga et al., 2006; Rost, 2003). The network feeds an input in the input layer that feeds the information to the hidden layer beneath it. In this hidden layer, nodes recode the information based upon weight for error balancing and pass it on to the next hidden layer (if present). If the information only flows forward, it is called a feed-forward network, while a feedback network also interacts between the nodes in a layer. Finally, the last hidden layer pushes the result to the output layer.

4.3 Support Vector Machines

A Support Vector Machine (SVM) is a supervised technique which was introduced by Vapnik in the late seventies (Vapnik, 1979). Conceptually, it non-linearly maps the input vectors to a very high-dimensional feature space. In this feature space, the SVM tries to construct a linear plane for classification (Cortes and Vapnik, 1995). This network was later extended to biological problems at the turn of the millennium (Hua and Sun, 2001; Cai et al., 2001) (Fig. 16).

![Illustration of data from input space, represented in higher dimensional feature space using kernels in the SVM.](image)

The SVM is a mathematical model that classifies a set of samples into two classes.
The linear separation of these two classes has a decision margin. The distance between a hyperplane, separating two classes, and the closest training samples to the decision surface is termed as the margin (Tarca et al., 2007). The dividing hyper plane can be of two-dimensional nature in its simplest form, though it is also possible to construct higher-dimensional planes. Kernels are developed with different mathematical representations of the data points onto these higher-dimensional spaces. Linear, polynomial, and radial basis function kernels are examples of common kernels used in a SVM to transform into higher dimensions and are used for better classification. The sample points that are crossed by these hyperplanes and are closest to maximum margin are called support vectors.

By choosing a suitable basis function, the SVM can transform a non-linear problem into a higher linear \textit{kernel} space. The linear model in the feature space then corresponds to the non-linear model in the input space (Cover, 1965; Sewell, 2011). The ease of training and ability to scale to high-dimensional data makes the SVM a strong tool in machine learning techniques (Markowitz, 2001). The major challenge though remains to identify the appropriate kernel for perfect classification of the classes for a given application (Burges, 1998). New kernels are constantly developed to get a better fitting of the hyperplane in a complex feature space for classification.
5 Paper Summary

5.1 Model quality assessment for membrane proteins based on support vector machines (Paper 1)

Our long-standing goal is to predict the 3D structure of membrane proteins with high accuracy to enable docking studies of proteins with unknown x-ray structures. One step in this direction is to be able to correctly identify and separate correct from less correct models from a set of plausible models also known as model quality assessment. Further, it is also important to identify which regions of a protein model that are correct and incorrect. Several studies have focused on these issues but so far all of them have excluded membrane proteins. This is unfortunate since there is a huge gap between the fraction of membrane proteins in the genome (25%) and the membrane proteins in the structural database (< 1%) making them an ideal target for computational approaches. In this study, we have improved and extended ProQ, one of the best methods for model quality assessment of water-soluble proteins, to membrane proteins.

ProQM was the first model quality assessment program that was solely developed for membrane proteins. It was developed based on support vector machines with a linear kernel. With a systematic analysis of the features, sequence conservation together with profile information showed the largest performance increase. This was expected as the sequence conservation and the profiles contain a large portion of information for the description of a protein structure.

Analysis of the results showed that the method was very good at detecting the residues with less than 3 Å deviation from the native structure residing in the membrane center. From the membrane core, the protein eventually becomes similar to globular domains. ProQM had a comparable performance to the current best globular MQAPs in those regions.

ProQM performed well while testing using the Skolnick set for GPCR proteins, over all 15 models. We also attempted decoy selection on a large set of ROSETTA-generated low-resolution membrane protein decoys, we performed decoy selection test. The combination of ProQM and ROSETTA showed seven-fold enrichment of the near-native decoys as the starting population of decoys.

5.2 Support vector machine based model quality assessment for globular proteins (Paper 2)

ProQ2 is a non-consensus based model quality assessment program for globular proteins. ProQ2 calculates structural properties based on the model and uses a SVM to predict the quality. The structural properties are similar to the ones used in the previous version of ProQ, but additional features that were shown to improve performance have been added.

The method uses a combination of features like evolutionary and multiple sequence alignment information along with other structural features to predict the local and global quality of a model. All features are calculated over a sliding sequence window and the local structural quality is calculated by S-score. A global score is obtained by summing up the local scores and normalizing the value by dividing by the target length. The program was trained by using a set of models generated in the 7th round of the CASP
experiment. Since it would be unfair to test the performance on the same data as used for training, scoring was instead performed on a similar dataset from CASP 8.

ProQ2 is significantly better than its predecessors at detecting high quality models, compared to the best single method in CASP8. The absolute quality assessment of the models at both local and global level is also improved. Combining ProQ2 with the consensus method, Pcons, we improved the model selection even further, surpassing the best servers presented in CASP 8.
6 Final thoughts

The protein structure prediction field has immense importance in drug design and structural and functional studies. It is today easy to build large number of protein models for any given sequence. Yet, we still lack the tools to consistently rank these models in order to select the best one. With more high throughput sequencing centers being created globally, the number of sequences with unknown structures will increase. As a consequence, the need for better model building and assessment methods will get increasingly important to bridge the gap between sequences with unknown structures vs experimentally deduced structures.

The contacts between residues are important for maintaining the native fold. Prediction of these residue-residue contacts has been a challenging problem in structural bioinformatics. Recently developed methods for accurate prediction of contacts will have implications in the field of fold recognition, ab initio protein folding, model quality assessment and de novo protein design (Jones et al., 2011). The successful methods extracted correlated mutation information from a MSA. In doing so, inherently, the methods showed a high tendency of false positives. This arose from the noise to the signal in the MSA because of the phylogenetic bias and the indirect coupling effect (Lapedes et al., 1999). The methods had a very low accuracy of around 20–40% in prediction (Graña et al., 2005; Hamilton et al., 2004; Fariselli et al., 2001; Jones et al., 2011). As recent as October 2011, there have been some ground breaking advances in the accuracy of these predictions (Morcos et al., 2011; Jones et al., 2011; Taylor et al., 2011). These new methods have shown very promising results with high accuracies. PSICOV has shown an average precision as high as 80% when predicting residues separated by 5–9 positions. The methods, though quite promising, still have room for improvement. Since all the methods use MSA, there is a need for better alignment tools to accurately align the tens of thousands of protein sequences being found by all the advances in high throughput sequencing. Also, these new methods are heavily dependent on the extraction of coevolutionary information from MSA. Hence, this is another aspect which needs further research.
7 Acknowledgment

First, I would like to thank Björn and Erik for believing in me. You have enriched my world with experiences, exposures and teachings. This was a dream come true and I have cherished every moment of it. Both of you inspire me each day to be better than yesterday.

To the Lindahl group, I feel grateful for being in the perfect environment for me to learn and grow. Sander, for the times of being there as a mentor and as a friend. Rossen, for sharing the common love of Indian food and music! Szilard, for being my officemate, with whom I have cherished many debates, for the past three years and being a company in the numerous late working nights. Torben, for all the parties, debates and venting we have shared. Sophie, for being there as a friend and a constant source of motivation.

Thank you Arne, for the numerous advices you gave me on bioinformatics. You have always been an “unofficial” mentor to me.

I am immensely indebted to my friends who have walked beside me, all through this journey. Linnea, you have been my rock in the most challenging times, more than once. Marcin, for the countless brainstorming sessions, along with our many ”politically inappropriate” debates! Walter, for the numerous karaoke sessions and being a dear friend.

It would be unfair on my part to not mention a few others who have been an integral part of my life in Stockholm. Thank you to the Skwark family for the numerous foody evenings and lovely company. Zak and Marion for always providing me a safe haven, away from the city buzz and for being a family to me. Maria, Lotta and Lollo for being more of a friend than a departmental component. I would also like to mention Stampen for giving me with the much needed distraction and relaxing evenings over this period of time - you shall always remain as my second home.

Last, but most importantly, my parents and brother. Without you three, I would not be where I am. The self-belief and standards that you have imbibed in me, shall always be my guide in life. The fire and will to stand against the general consensus has always pushed to me to go beyond. I will always remember to ‘never give up’... because you guys never gave up on me! This thesis I dedicate to you...

Arjun Ray, Stockholm, March 2012
List of Figures

1. A DNA codon consisting of three bases on the backbone of phosphate and sugar. ........................................ 1
2. The protein synthesis pathway. The DNA is expressed into the RNA through transcription process. The RNA is then translated into a protein. .................................................. 2
3. The mRNA copies in a complementary manner from the DNA. ............... 3
4. The tRNA binds with the amino acids. ................................................ 3
5. The first tRNA links to the left-most codon of the mRNA strand in the ribosome. ........................................... 3
6. The basic structure of an amino acid contains a carboxylic group, an amino group and a side chain. ................. 4
9. Structure of a mutant KcsA potassium channel with its secondary structures. PDBID: 1ZWI. .................................................. 11
10. The energy landscape of protein folding. .............................................. 15
11. In the last decade, the cost of DNA sequencing has fallen to a hundred-thousandth of its initial cost (Carr, 2010). ................................. 16
12. Illustration of a multiple sequence alignment result for the first 40 residues of hemoglobin using ClustalW (Thompson et al., 1994). ............. 18
13. A: Models with less than 30% sequence identity to the target frequently have substantial alignment errors and a suboptimal template selection. B: Models with 30-50% sequence identity to the target frequently have alignment errors in the non-conserved segments of the target protein, structural variation in templates, and incorrect reconstruction of the loops. C: Models having higher than 50% sequence identity to the target typically have the correct fold and the alignments tends to be mainly correct. .... 21
14. Flowchart of homology modeling. Courtesy of Dr. Björn Wallner. .... 21
15. Predicted as well as true deviation for an early model of the β2-adrenergic receptor. This GPCR model by Skolnick (Zhang et al., 2006) was published before any experimental structure was available, but ProQM correctly identifies regions where the model was later found to deviate from the X-ray structure (Ray et al., 2010). ................................................. 24
16. Illustration of data from input space, represented in higher dimensional feature space using kernels in the SVM. .................. 26
References


