Approaches to Protein Structure Prediction and Their Applications

S. Selvaraj*
Department of Physics
Bharathidasan University
Tiruchirappalli 620 024
* e-mail:(sel_emi@yahoo.co.uk; selva@physics.bdu.ac.in)

Introduction
Proteins are one of the major biological macromolecules performing a variety functions such as enzymatic catalysis, transport, regulation of metabolism, nerve conduction, immune response etc. The three-dimensional structure of a protein is responsible for its function. In this paper I wish to give an overview of the need for protein structure prediction, the different approaches available as of now and their applications and limitations. Special mention is made of the utility of homology derived models in the drug discovery process.

Sequence-Structure Gap and the Need for Structure Prediction

With the advent of recombinant DNA technology it has become possible to determine the amino acid sequences of proteins quite rapidly. However, determining the three-dimensional structure of proteins is a time consuming task and hence there exists a vast gap between the number of proteins of known amino acid sequence and that of known structures. This is called as the sequence-structure gap. As the knowledge of the 3-D structure of a protein is very essential to understand its function, it is imperative to develop techniques to predict the structure of a protein from its amino acid sequence.

Basis for Structure Prediction:

The classic experiments carried out by C.B. Afinson in the 60’s on the enzyme ribonuclease led to the conclusion that the information to specify the 3-D structure of a protein resides in its amino acid sequence. Within the cell a newly synthesized protein chain spontaneously folds into the compact globular structure to perform its function. Thus nature has an algorithm to fold proteins to their native structures. Efforts have been directed for the
past four decades to discover nature’s algorithm and computational methods have been developed to predict the structure of proteins from their sequences.

Approaches to Structure Prediction

Prediction of protein structures can be classified into two major categories viz. (i) Prediction of secondary structure and (ii) Prediction of tertiary (3-D) structure. Prediction of secondary structure of proteins attempts to locate segments of the polypeptide chain adopting the α-helical or β-strand structure. Regions that are devoid of these regular secondary structural elements are considered to adopt coil conformation. In tertiary structure prediction, one attempts to predict the three-dimensional structure of a protein or the native structure. While so far this has remained an elusive goal, different methods have been developed to press forward to the attainment of this goal. Table 1 provides a list of web-servers useful in protein structure prediction.

Secondary structure Prediction

Since the early 1970s, various predictive algorithms have been developed based on the knowledge of conformations of amino acid residues as observed in the crystal structure of proteins. Amongst all the methods, the Chou-Fasman (CF) and the Garnier-Osguthorpe-Robson methods (GOR) methods have been widely employed. In recent years elaborate machine learning methods such as neural networks, nearest-neighbor techniques and hidden Markov models etc. have been developed. A consensus prediction constructed from the output of several methods seems to result in improvement of prediction. For a given sequence if a number of homologous sequences are available in the sequence database, these may be aligned to give a multiple alignment. This multiple alignment can be used as input for secondary structure prediction which gives more reliable results.

Application of secondary structure prediction

Though secondary structure prediction of proteins is not an end in itself, it supports other prediction problems such as identification of remote homologs i.e. discrimination of related and unrelated proteins in the range of 10 – 30 % sequence identity. It also helps in fold recognition and structural clustering.
Tertiary (3-D) Structure Prediction

Presently three approaches are followed for the prediction of the tertiary structure of proteins. These are 1) Homology modelling (Comparative Modelling), 2) Threading and 3) Ab initio structure prediction.

1. Homology Modelling

This is the simplest and most reliable approach. The observation that proteins with similar sequences tend to fold into similar structures forms the basis for this method. It has been observed that even proteins with 25% sequence identity fold into similar structures. This method does not work for remote homologs (< 25% pairwise identity). The algorithm for homology modelling may be briefly stated as: given a query sequence $Q$, and a sequence database of known protein structures, find a protein $P$ such that $P$ has high sequence similarity to $Q$ and return $P'$’s structure as an approximation to $Q'$ structure. The following are the main steps in homology modelling:

1) Finding known structures (templates) related to the query sequence whose structure has to be modeled
2) Aligning the query sequence to the template(s)
3) Constructing variable side-chains and main-chains (loops, insertions and deletions) and
4) Model refinement, assessing the model(s) built and selecting the most native conformations.

2. Threading

Threading is a method for fold recognition. This is used for sequences with sequence identity $\leq$ 30%. In this approach, given a sequence and the set of folds available in the Protein Data Bank (PDB) the aim is to see if the sequence can adopt one of the folds of known structure. This method takes advantage of the knowledge of existing structures and the principles by which they are stabilized. Fold assignment and alignment are achieved by threading the sequence through each of the structures in a library of all known folds.
3. *Ab initio* (de novo) structure prediction

While homology modelling and threading requires knowledge of known structures, *ab initio* structure prediction has no such limitations. It starts with the assumption that the native structure of a protein is at the global free energy minimum. A large scale search of conformational space for protein structures that are particularly low in free energy for the given amino acid sequence are carried out. In recent years a particularly successful method called **Rosetta** has been developed by Baker and colleagues (Simons et al, 1997). This method has assimilated information obtained from known structures and is based on a picture of protein folding in which short segments of the protein chain flicker between different local structures consistent with their local sequence, and folding to the native state occurs when these local segments are oriented such that low free energy interactions are made throughout the protein.

**Applications and limitations of methods for structure prediction**

Each of the above methods discussed provide structural details to different extent. While homology modelling can provide atomic level details of the target protein, threading can help only to judge the fold of the protein. Baker and Sali (2001) have discussed the accuracy and application of protein structure models with examples. High and medium level homology models with sequence identity $> 30\%$ are useful in refining functional prediction such as ligand binding. Low accuracy models of many of the ribosomal proteins were helpful in building the molecular model for whole yeast ribosome. Folds predicted by threading could be used in supporting site-directed mutagenesis experiments, designing of stable crystallizable variants and in refining NMR structures. The accuracy and applicability of models produced by *ab initio* methods are in general of lower accuracy compared to models obtained from either homology modelling or threading. These are useful in predicting functional relationships from structural similarity and for identification of patches of conserved surface residues.

**Critical Assessment of Structure Prediction (CASP)**

Judging techniques for predicting structures requires blind tests. A decade ago the Critical Assessment of methods of protein Structure Prediction (CASP) meetings were initiated by J. Moult and others. These meetings are held once in two years and provide a benchmark for the evaluation of protein structure prediction models. For this purpose, crystallographers and NMR
Spectroscopists in the process of determining a protein structure are invited to publish the amino acid sequence several months before the expected date of completion and to keep the results secret. Predictors submit models based on sequence information. The predictions and experimental results are compared. These meetings help to focus on the progress made in the structure prediction methods and areas that need further refinement.

The CASP web-site can be visited at: [http://predictioncenter.llnl.gov](http://predictioncenter.llnl.gov)

Application of homology models in drug discovery

Homology models can be used at various stages of drug discovery process. Homology models help in the process of assessing the drugability of a given protein. The availability of computational methods to identify binding pockets in the structures of proteins and a large number of homology models has opened new possibilities for identifying proteins that can be targeted by a drug. Further homology models help in structure-guided design of mutagenesis experiments that in turn enables one for target validation studies. Numerous examples are reported in the literature where protein homology models have supported the discovery and the optimization of lead compounds with respect to potency and selectivity (Hillisc et al. 2004).

**Conclusions**

This chapter has provided a brief overview of protein structure prediction methods and the basis of each method. It has also highlighted some applications and limitations of the various methods. In the absence of experimental structural information homology based models are very valuable in the drug discovery process. It must be borne in mind that structure prediction is an area of continuous research and development. It is suggested that the results from related methods be used to get a consensus structure which in general can be expected to be nearer to the actual structure.
REFERENCES


Table 1. List of some structure prediction servers

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Servers for Secondary Structure Prediction</strong></td>
<td></td>
</tr>
<tr>
<td>PHD</td>
<td><a href="http://cubic.bioc.columbia.edu/predictprotein">http://cubic.bioc.columbia.edu/predictprotein</a></td>
</tr>
<tr>
<td>PSI-PRED</td>
<td><a href="http://insulin.brunel.ac.uk/psipred">http://insulin.brunel.ac.uk/psipred</a></td>
</tr>
<tr>
<td>JPRED</td>
<td><a href="http://jura.ebi.ac.uk:8888">http://jura.ebi.ac.uk:8888</a></td>
</tr>
<tr>
<td>NPSA</td>
<td><a href="http://npsa-pbil.ibcp.fr">http://npsa-pbil.ibcp.fr</a></td>
</tr>
<tr>
<td><strong>Servers for Homology (Comparative) Modelling</strong></td>
<td></td>
</tr>
<tr>
<td>RAMP</td>
<td><a href="http://software.compbio.washington.edu/ramp/">http://software.compbio.washington.edu/ramp/</a></td>
</tr>
<tr>
<td>SWISSMODEL</td>
<td><a href="http://swissmodel.expasy.org">http://swissmodel.expasy.org</a></td>
</tr>
<tr>
<td><strong>Servers for Threading (Fold Recognition)</strong></td>
<td></td>
</tr>
<tr>
<td>3-D PSSM</td>
<td><a href="http://www.sbg.bio.ic.ac.uk/~3dpssm">http://www.sbg.bio.ic.ac.uk/~3dpssm</a></td>
</tr>
<tr>
<td>UCLA-DOE Fold Server</td>
<td><a href="http://fold.doe-mbi.ucla.edu/">http://fold.doe-mbi.ucla.edu/</a></td>
</tr>
</tbody>
</table>