International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 6 Issue 8|| August 2017 || PP. 33-37

Structural Determination of Proteins Using Computation

[*]Dr. Suranjana Chattopadhyay
Assistant Professor in Chemistry, Maharaja Manindra Chandra College
Corresponding Author: Dr. Suranjana Chattopadhyay

Date of Submission: 11-07-2017 Date of acceptance: 05-08-2017

I. INTRODUCTION

Rapid advancements in the post-genomic era along with the introduction of novel sequencing technologies provided an even platform for the researchers around the world to sequence new protein and nucleotide sequences in a faster and efficient manner. The fundamental biological concept of "Sequence implies the Structure and Structure implies the Function" deciphers that this increase in the amount of sequence knowledge does not reflects any biological significance until the structure of the protein is identified. Criticality the biological function of a protein is totally dependent on its native 3D structure.

Why?-

• A full understanding of a molecular system comes from careful examination of the sequence-structure-function triad

• Below 30% protein sequence identity detection of a homologous relationship is not guaranteed by sequence alone

• Structure is much more conserved than sequence

Frequently applied experimental protein structure determination techniques viz. XRD or NMR are quite accurate. But the computational models on the other hand are distinct as well as comprehensive tools to predict a wide landscape of proteins taxonomically.

Protein Predicting Strategies

If a protein of known tertiary structure shares at least 30% of its sequence with a potential homolog of undetermined structure, comparative methods that overlay the putative unknown structure with the known can be utilized to predict the likely structure of the unknown. However, below this threshold three other classes of strategy are used to determine possible structure from an initial model: **ab initio protein prediction, fold recognition, and threading**.

1. **Ab Initio Methods:** In ab initio methods, an initial effort to elucidate secondary structures (alpha helix, beta sheet, beta turn, etc.) from primary structure is made by utilization of physicochemical parameters and neural net algorithms. From that point, algorithms predict tertiary folding. One drawback to this strategy is that it is not yet capable of incorporating the locations and orientation of amino acid side chains.
2. **Fold Prediction:** In fold recognition strategies, a prediction of secondary structure is first made and then compared to either a library of known protein folds, such as CATH or SCOP, or what is known as a "periodic table" of possible secondary structure forms. A confidence score is then assigned to likely matches.
3. **Threading:** In threading strategies, the fold recognition technique is expanded further. In this process, empirically based energy functions for the interaction of residue pairs are used to place the unknown protein onto a putative backbone as a best fit, accommodating gaps where appropriate. The best interactions are then accentuated in order to discriminate amongst potential decoys and to predict the most likely conformation.

The goal of both fold and threading strategies is to ascertain whether a fold in an unknown protein is similar to a domain in a known one deposited in a database, such as the protein databank (PDB). This is in contrast to de novo (ab initio) methods where structure is determined using a physics-base approach in lieu of comparing folds in the protein to structures in a data base.

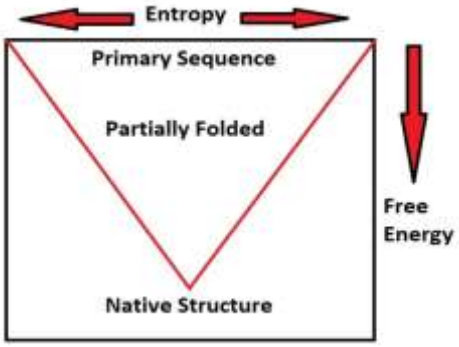

Fig.1. Correctly folded protein conformations (native structures) have lower free energies than partially folded or primary structures. Computers search for these conformations because they indicate correct folding

The structure-function relationship is even more complex than the relationship between sequence and structure (and not as well understood)

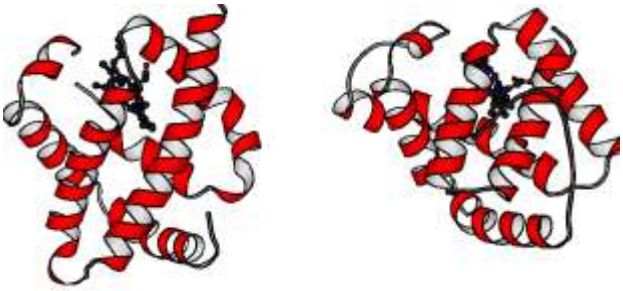

Fig.2. The globin fold is resilient to amino acid changes. V. stercoraria (bacterial) hemoglobin (left) and P. marinus (eukaryotic) hemoglobin (right) share just 8% sequence identity, but their overall fold and function is identical.

Improved sequence to structure alignment residuals with better energy functions for evaluating the fit may allow precise fold recognition and alignment in threading studies. Moreover, refinement of the predicted model by the Molecular Dynamics simulations can prove to be major breakthrough in adjustments to side chain stereochemistry, backbone conformation and model correction (Zhao et al., 2013; Nygaard et al., 2013). Protein structure refinement during CASP11 by the Feig group was described. Molecular Dynamics simulations were used in combination with an improved selection and averaging protocol. The internal structure of each of the molecular fragments is treated realistically, while there is no interaction between different molecular fragments to avoid unphysical steric clashes. The information from all the molecular fragments is exploited simultaneously to constrain the backbone to refine a three-dimensional model of the conformational state of the protein. On average, modest refinement was achieved with some targets improved significantly.

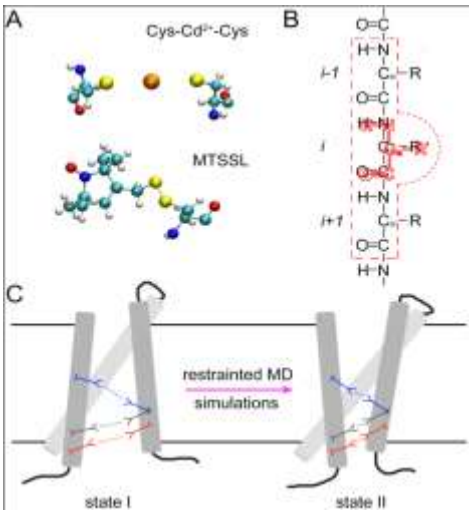

Fig.3. Restrained Molecular Dynamics

However despite of significant progress researchers in their ability to further increase the quality of the models to the experimental level have been delimited by several challenges and shortcomings. The computational models often represent only fractions of the full-length of desired protein leaving behind the unresolved questions in template-based modeling to combine information from multiple templates, viz., different structural domains, into larger complex assemblies. The development of consistent, accurate and progressive methods for improvement of models by shifting the coordinates parallel to the native state is one of the burning issues (MacCallum et al., 2011]; Wass et al., 2011). To some extent the largest possibilities in the escalation of the models came from more experimentally determined structures which allow better conceivable templates for the targets.

Protein structure prediction

It is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary and tertiary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction).

The practical role of protein structure prediction is now more important than ever. Massive amounts of protein sequence data are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. The protein structure prediction remains an extremely difficult and unresolved undertaking. The two main problems are calculation of protein free energy and finding the global minimum of this energy. A protein structure prediction method must explore the space of possible protein structures which is astronomically large.

Energy- and fragment-based methods

Ab initio- or *de novo*- protein modelling methods seek to build three-dimensional protein models "from scratch", i.e., based on physical principles rather than (directly) on previously solved structures. There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search possible solutions (i.e., global optimization of a suitable energy function). These procedures tend to require vast computational resources, and have thus only been carried out for tiny proteins. To predict protein structure *de novo* for larger proteins will require better algorithms and larger computational resources like those afforded by either powerful supercomputers (such as Blue Gene or MDGRAPE-3) or distributed computing (such as Folding@home, the Human Proteome Folding Project and Rosetta@Home). Although these computational barriers are vast, the potential benefits of structural genomics (by predicted or experimental methods) make *ab initio* structure prediction an active research field.

As of 2009, a 50-residue protein could be simulated atom-by-atom on a supercomputer for 1 millisecond. As of 2012, comparable stable-state sampling could be done on a standard desktop with a new graphics card and more sophisticated algorithms. A much larger simulation timescales can be achieved using coarse-grained modeling.

Evolutionary co-variation to predict 3D contacts

As sequencing became more commonplace in the 1990s several groups used protein sequence alignments to predict correlated mutations and it was hoped that these coevolved residues could be used to predict tertiary structure (using the analogy to distance constraints from experimental procedures such as NMR). The assumption is when single residue mutations are slightly deleterious, compensatory mutations may occur to re-stabilize residue-residue interactions. This early work used what are known as *local* methods to calculate correlated mutations from protein sequences, but suffered from indirect false correlations which result from treating each pair of residues as independent of all other pairs.

In 2011, a different, and this time *global* statistical approach, demonstrated that predicted coevolved residues were sufficient to predict the 3D fold of a protein, providing there are enough sequences available (>1,000 homologous sequences are needed). The method, **EVfold**, uses no homology modeling, threading or 3D structure fragments and can be run on a standard personal computer even for proteins with hundreds of residues. The accuracy of the contacts predicted using this and related approaches has now been demonstrated on many known structures and contact maps, including the prediction of experimentally unsolved transmembrane proteins.

From a structural perspective, a protein is an ordered linear chain of building blocks known as amino acid residues. Each protein is defined by its unique sequence of amino acids. This sequence causes the protein to fold into a particular three-dimensional shape. Predicting the folded structure of a protein only from its amino acid sequence remains a challenging problem in mathematical optimization (Lander and Waterman, 1999). The

challenge arises due to the combinatorial explosion of plausible shapes each of which represent a local minimum of an intricate non-convex function of which the global minimum is sought. In nature, proteins typically present 50 to 500 amino acid residues. The books by Lesk (Lesk, 2002) and Tramontano (Tramontano, 2006) present elegant, comprehensive overviews of protein structure. In nature there are 20 distinct proteinogenic amino acids, each one with its own chemical properties (including size, charge, polarity, hydrophobicity, i.e. the tendency to avoid water packing) (Lodish et al., 1990; Lehninger et al., 2005). Depending on the polarity of the side-chain, amino acids vary in their hydrophilic or hydrophobic character. The importance of the physical properties of the side-chains comes from the influence they have on the amino acid residues interactions in the 3-D structure. The distributions of the hydrophilic and hydrophobic amino acids are important to determine the tertiary structure of the polypeptide.

A peptide is a molecule composed of two or more amino acid residues chained by a chemical bond called the peptide bond. This peptide bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, thereby releasing a water molecule. Two or more linked amino acid residues are referred to as a peptide, and larger peptides are generally referred to as polypeptides or proteins (Creighton, 1990; Lesk, 2002).

Structure determination of proteins using ^{1}H NMR and Conformational Analysis

First of all the amino acid residues in the primary sequence of the polypeptide chain are identified typically according to their chemical shifts which produce a typical pattern in the **Total Correlation Spectroscopy experiment (TOCSY).**As the name suggests TOCSY experiment presents us the total correlation picture of the spin systems present in the polypeptide.

According to all the chemically non-equivalent protons present in a particular amino acid, TOCSY draws a total picture of correlation among them by subsequent 3-bond coupling which is carried through the entire side chain of the amino acids and through the polypeptide backbone at the same time. It can be compared to a relay experiment whose final outcome is to move along the primary sequence of the polypeptide identifying the various types of amino acids present.

Next stage is to sequence assign the polypeptide chain because an amino acid may occur more than once in the series of amino acid residues, and the exact order of occurrence of each residue must be known in order to determine the nature of the polypeptide or protein. A strategy that is used here is called the **NOE walk.** The □-hydrogen of **an ith amino acid residue couples with the amide proton of residue i+1**. Hence once a particular amino acid's spin system is assigned (from TOCSY), using the □ N(i,i+1) NOE cross peak (from NOESY) the next amino acid in the sequence can be identified. This walk is followed until the last amino acid in the sequence is reached.

After the NOESY generates all possible proton pairs that are close in distance, a table of inter and intra residual proton-proton distances are compiled. If there are sufficiently large number of inter-proton short distances, one can go further and build a model structure using these as constraints in modeling softwares like the **DYANA, CYANA** which now replaces DYANA and allows for automated NMR structure calculation. Given a sufficiently complete list of assigned chemical shifts and one or several NOESY spectra, the assignment of the NOESY cross-peaks and the three-dimensional structure of the protein in solution can be calculated automatically.

This way a conformational analysis and subsequent protein structure design may be done from NMR studies coupled with modeling. Apart from the distance restraints, another major constraint that may be used to facilitate the modeling as stated above is the **three bond coupling constant** values between the amide protons and the alpha hydrogens. $^{3}J_{N\square}\square$ can provide valuable information about the dihedral angle □ □ □he peptide bond (C-N) has a double bond and is not allowed rotation of the molecule around this bond. The rotation is only permitted around the bonds N-C□ and C□-C. These bonds are known as PHI □ □and PSI □ angles, respectively, and are free to rotate. This freedom is mostly responsible for the conformation adopted by the polypeptide backbone. As a consequence, the possible conformation of a given polypeptide is quite limited and depends on the amino acid chemical properties. The peptide bond itself tends to be planar, with two allowed states: trans, ω 180◦ (usually) and cis, ω 0◦ (rarely) (Branden and Tooze, 1998; Lesk, 2002).

Three-dimensional structure of a small cationic protein from Marine Turtle was determined by distance geometry and simulated annealing. Fig.4. (A) shows energy-minimized structure of the protein. Fig.4. (C) also shows the electrostatic potential of the turtle egg white protein. The protein shows clustering of positive charges followed by relative electro neutrality.

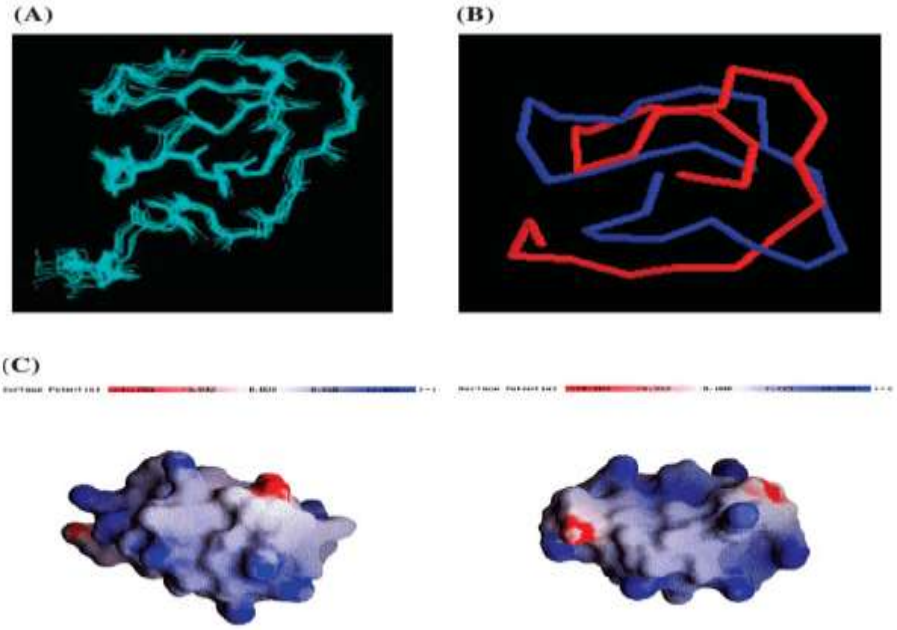

Fig.4.(A): The cluster of 10 energy minimized structures of the egg white protein (B): The egg white protein(residues 3-26, red) superimposed on hbd-3 (residues 18-41, blue) (C): Electrostatic Potential surface of the protein (generated by GRASP).

(Ref: **Small Cationic Protein From a Marine Turtle Has □ □Defensin-Like Fold and Antibacterial and Antiviral Activity. Chattopadhyay, S., Sinha, N. K., Banerjee, S., Roy, S., Chattopadhyay, D., and Roy, S.** *PROTEINS: Structure, Function, and Bioinformatics, (2006) 64,524–531)*

II. CONCLUSION

The intention of the protein structure prediction problem is to find out the structure from a given amino acid sequence. In this article gone all the way, through many of the evolutionary algorithms, used to anticipate the structure, tools are listed out to find out a possibly precise solution to a protein structure by computational methods from the experimental data available. There are multifarious other ways of approaching a protein 3-D structure including Comparative Protein Modeling strategies like **homology modeling** process. Based on the protein database one can also easily find the particular protein id and all those information about the specific protein.

REFERENCES:

[1] Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. Gergely Csaba, Fabian Birzele and Ralf Zimmer BMC Structural Biology 2009
[2] Kini, R.M., Evans, H.J., Molecular modelling of proteins : a strategy for energy minimization by molecular mechanics in the AMBER force field. J Mol Struct Dyn 1991 Dec; 9(3), 475-88
[3] Zhao, G., Perilla, J.R., Yufenyuy, E.L., Meng, X., Chen, B., Ning, J., et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature. 2013; 497, 643–6
[4] MacCallum, J.L., Perez, A., Schnieders, M.J., Hua, L., Jacobson, M.P., Dill, K.A. Assessment of protein structure refinement in CASP9. Proteins. 2011;79, 74–90.
[5] Introduction to Protein Structure- 17 Dec 1998**,**by Carl Ivar Branden , John Tooze
[6] Introduction to Bioinformatics by Arthur M. Lesk (2002-02-14**)** by Arthur M. Lesk
[7] Moult, J.; et al. A large-scale experiment to assess protein structure prediction methods. Proteins. (1995). 23 (3), ii–iv.
[8] Güntert, P., Mumenthaler, C. & Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. . (1997) 273, 283-298.
[9] Güntert, P. Automated NMR protein structure calculation with CYANA. Meth. Mol. Biol. (2004). 278, 353-378.
[10] Chattopadhyay, S., Sinha, N. K., Banerjee, S., Roy, S., Chattopadhyay, D., and Roy, S. Small Cationic Protein From a Marine Turtle Has □ □Defensin-Like Fold and Antibacterial and Antiviral Activity. PROTEINS: Structure, Function, and Bioinformatics, (2006) 64,524–531

International Journal of Engineering Science Invention (IJESI) is UGC approved Journal with Sl. No. 3822, Journal no. 43302.

*Dr. Suranjana Chattopadhyay " Structural Determination of Proteins Using Computation ." International Journal of Engineering Science Invention (IJESI) 6.8 (2017): 33-37.