International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-3, Issue-10)

Research Article | **October 2014**

OverView of Protein Structure Prediction in Bioinformatics

N. Deepak Kumar*
Dept. of CSE& Svuniversity,
Tirupati, India

Dr. A. Ramamohan Reddy
Dept. of CSE & Svuniversity,
Tirupati, India

Abstract—

Protein structure prediction from a sequence is one of the high focus problems for researchers.This is a very useful application of bioinformatics as the experimental techniques like X-ray crystallography are time consuming.The fundamental issue is how can we predict the 3-D shape of a protein from its amino acid sequence. In this paper we will learn how to predict protein structure and function based on the amino acid sequence.

Keywords— Bioinformatics,Protein,Primary,Secondary,Tertiary,Quaternary.

I. INTRODUCTION

Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. (Of course, there are exceptions, such as the bovine spongiform encephalopathy – a.k.a. Mad Cow Disease – prion.) Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as one of *secondary*, *tertiary* and *quaternary* structure. A viable general solution to such predictions remains an open problem. Most efforts have so far been directed towards heuristics that work most of the time[1,2].One of the key ideas in bioinformatics is the notion of homology. In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene *A*, whose function is known, is homologous to the sequence of gene *B*, whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably[4,5].One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes (leghemoglobin). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes[11].Other techniques for predicting protein structure include protein threading and *de novo* (from scratch) physics-based modeling.Tens of thousands of three-dimensional protein structures have been determined by X-ray crystallography and protein nuclear magnetic resonance spectroscopy (protein NMR) and a central question in structural bioinformatics is whether it is practical to predict possible protein–protein interactions only based on these 3D shapes, without performing protein–protein interaction experiments. A variety of methods have been developed to tackle the protein–protein docking problem, though it seems that there is still much work to be done in this field[14].Other interactions encountered in the field include Protein–ligand (including drug) and protein–peptide. Molecular dynamic simulation of movement of atoms about rotatable bonds is the fundamental principle behind computational algorithms, termed docking algorithms, for studying molecular interactions[14].

II. METHODOLOGY

The more commonly used terms for evolutionary and structural relationships among proteins are listed below. Many additional terms are used for various kinds of structural features found in proteins. Descriptions of such terms may be found at the CATH Web site the Structural Classification of Proteins (SCOP) Web site and a Glaxo-Wellcome tutorial on the Swiss bioinformatics Expasy Web site.

- **active site**
  a localized combination of amino acid side groups within the tertiary (three-dimensional) or quaternary (protein subunit) structure that can interact with a chemically specific substrate and that provides the protein with biological activity. Proteins of very different amino acid sequences may fold into a structure that produces the same active site.
- **architecture**
  the relative orientations of secondary structures in a three-dimensional structure without regard to whether or not they share a similar loop structure.
- **fold**
  a type of architecture that also has a conserved loop structure.
- **blocks**
  a conserved amino acid sequence pattern in a family of proteins. The pattern includes a series of possible matches at each position in the rep- resented sequences, but there are not any inserted or deleted positions in the pattern or in the

sequences. By way of contrast, sequence profiles are a type of scoring matrix that represents a similar set of patterns that includes insertions and deletions.

- **class**
  a term used to classify protein domains according to their secondary structural content and organization. Four classes were originally recognized by Levitt and Chothia (1976), and several others have been added in the SCOP database. Three classes are given in the CATH database: mainly-α, mainly-β, and α–β, with the α–β class including both alternating α /B and α+β structures.
- **core**
  the portion of a folded protein molecule that comprises the hydrophobic interior of α-helices and β-sheets. The compact structure brings together side groups of amino acids into close enough proximity so that they can interact. When comparing protein structures, as in the SCOP database, core is the region common to most of the structures that share a common fold or that are in the same superfamily. In structure prediction, core is sometimes defined as the arrangement of secondary structures that is likely to be conserved during evolutionary change.
- **domain (sequence context)**
  a segment of a polypeptide chain that can fold into a three-dimensional structure irrespective of the presence of other segments of the chain. The separate domains of a given protein may interact extensively or may be joined only by a length of polypeptide chain. A protein with several domains may use these domains for functional interactions with different molecules.
- **family (sequence context)**
  a group of proteins of similar biochemical function that are more than 50% identical when aligned. This same cutoff is still used by the Protein Information Resource (PIR). A protein family comprises proteins with the same function in different organisms (orthologous sequences) but may also include proteins in the same organism (paralogous sequences) derived from gene duplication and rearrangements. If a multiple sequence alignment of a protein family reveals a common level of similarity throughout the lengths of the proteins, PIR refers to the family as a homeomorphic family. The aligned region is referred to as a homeomorphic domain, and this region may comprise several smaller homology domains that are shared with other families. Families may be further subdivided into subfamilies or grouped into superfamilies based on respective higher or lower levels of sequence similarity. The SCOP database reports 1296 families and the CATH database (version 1.7 beta), reports 1846 families.
  When the sequences of proteins with the same function are examined in greater detail, some are found to share high sequence similarity. They are obviously members of the same family by the above criteria. However, others are found that have very little, or even insignificant, sequence similarity with other family members. In such cases, the family relationship between two distant family members A and C can often be demonstrated by finding an additional family member B that shares significant similarity with both A and C. Thus, B provides a connecting link between A and C. Another approach is to examine distant alignments for highly conserved matches.
  At a level of identity of 50%, proteins are likely to have the same three-dimensional structure, and the identical atoms in the sequence alignment will also superimpose within approximately 1 Å in the structural model. Thus, if the structure of one member of a family is known, a reliable prediction may be made for a second member of the family, and the higher the identity level, the more reliable the prediction. Protein structural modeling can be performed by examining how well the amino acid substitutions fit into the core of the three-dimensional structure.
- **family (structural context)**
  as used in the FSSP database (Families of structurally similar proteins) and the DALI/FSSP Web site, two structures that have a significant level of structural similarity but not necessarily significant sequence similarity.
- **fold**
  similar to structural motif, includes a larger combination of secondary structural units in the same configuration. Thus, proteins sharing the same fold have the same combination of secondary structures that are connected by similar loops. An example is the Rossman fold comprising several alternating α helices and parallel β strands. In the SCOP, CATH, and FSSP databases, the known protein structures have been classified into hierarchical levels of structural complexity with the fold as a basic level of classification.
- **homologous domain (sequence context)**
  an extended sequence pattern, generally found by sequence alignment methods, that indicates a common evolutionary origin among the aligned sequences. A homology domain is generally longer than motifs. The domain may include all of a given protein sequence or only a portion of the sequence. Some domains are complex and made up of several smaller homology domains that became joined to form a larger one during evolution. A domain that covers an entire sequence is called the homeomorphic domain by PIR (Protein Information Resource).
- **module**
  a region of conserved amino acid patterns comprising one or more motifs and considered to be a fundamental unit of structure or function. The presence of a module has also been used to classify proteins into families.
- **motif (sequence context)**
  a conserved pattern of amino acids that is found in two or more proteins. In the Prosite catalog, a motif is an amino acid pattern that is found in a group of proteins that have a similar biochemical activity, and that often is near the active site of the protein. Examples of sequence motif databases are the Prosite catalog (http://www.expasy.ch/prosite) and the Stanford Motifs Database (http://dna.stanford.edu/emotif/).

- **motif (structural context)**
  a combination of several secondary structural elements produced by the folding of adjacent sections of the polypeptide chain into a specific three-dimensional configuration. An example is the helix-loop-helix motif. Structural motifs are also referred to as supersecondary structures and folds.
- **position-specific scoring matrix** (sequence context, also known as weight or scoring matrix)
  represents a conserved region in a multiple sequence alignment with no gaps. Each matrix column represents the variation found in one column of the multiple sequence alignment.
- **Position-specific scoring matrix**—3D (structural context) represents the amino acid variation found in an alignment of proteins that fall into the same structural class. Matrix columns represent the amino acid variation found at one amino acid position in the aligned structures.

III. RESULT

The above techniques for protein function prediction use the strategy of predicting the protein functions by classification into putative function groups.They usually fail to predict specific protein function.Expert system have been built for prediction at a protein level.One of such expert system is **GeneQuiz**(http://bric.postech.ac.kr/seminar/kjh/GeneQuiz_biowave/tsld001.htm).

IV. CONCLUSION

In this paper ,we have learnt protein structure prediction that is a very useful and important application in bioinformatics.If the amino acid sequence of a protein is known,one can predict the protein structure,its properties and functions,but the situation is compounded due to protein folding problem.A number of protein identification and characterization tools are available.However predicting the structure and functions of transmembrance helices,a special class of protein that include GPCRs,is much needed,as they are important for therapeutic interactions.Although excellent tools and computational methods are available,none of the techniques is fool proof and the area remains a very exciting one for the researchers.

REFERENCES

[1] Hogeweg, P. (2011). "The Roots of Bioinformatics in Theoretical Biology". In Searls, David B. *PLoS Computational Biology* 7 (3): e1002021. Bibcode:2011PLSCB...7E0020H. doi:10.1371/journal.pcbi.1002021. PMC 3068925. PMID 21483479. edit

[2] Hesper B, Hogeweg P (1970). *Bioinformatica: een werkconcept* 1 (6). Kameleon. pp. 28–29.

[3] Hogeweg, P. (1978). "Simulating the growth of cellular forms". *Simulation* 31 (3): 90–96. doi:10.1177/003754977803100305. edit

[4] Moody, Glyn (2004). *Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business*. ISBN 978-0-471-32788-2.

[5] Dayhoff, M.O. (1966) Atlas of protein sequence and structure. National Biomedical Research Foundation, 215 pp.

[6] Eck RV, Dayhoff MO. (1966) Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences. Science. 1966 Apr 15;152(3720):363-366

[7] Johnson, George; Tai Te Wua (January 2000). "Kabat Database and its applications: 30 years after the first variability plot". *Nucleic Acids Res* 28 (1): 214–218. doi:10.1093/nar/28.1.214. PMC 102431. PMID 10592229.

[8] Sanger, F.; Air, G. M.; Barrell, B. G.; Brown, N. L.; Coulson, A. R.; Fiddes, J. C.; Hutchison, C. A.; Slocombe, P. M.; Smith, M. (1977). "Nucleotide sequence of bacteriophage φX174 DNA". *Nature* 265 (5596): 687–95. Bibcode:1977Natur.265..687S. doi:10.1038/265687a0. PMID 870828.

[9] Attwood TK, Gisel A, Eriksson N-E, Bongcam-Rudloff E (2011). "Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective". *Bioinformatics – Trends and Methodologies*. InTech. Retrieved 8 Jan 2012.

[10] Froimowitz M, Fasman GD (1974). "Prediction of the secondary structure of proteins using the helix-coil transition theory". *Macromolecules* 7 (5): 583–9. doi:10.1021/ma60041a009. PMID 4371089

[11] Dor O, Zhou Y (2006). "Achieving 80% tenfold cross-validated accuracy for secondary structure prediction by large-scale training". *Proteins* 66 (4): 838–45. doi:10.1002/prot.21298. PMID 17177203.

[12] Zhong L, Johnson WC Jr (1992). "Environment affects amino acid preference for secondary structure". *Proc Natl Acad Sci USA* 89 (10): 4462–5. doi:10.1073/pnas.89.10.4462. PMC 49102. PMID 1584778.

[13] Macdonald JR, Johnson WC Jr (2001). "Environmental features are important in determining protein secondary structure". *Protein Sci.* 10 (6): 1172–7. doi:10.1110/ps.420101. PMC 2374018. PMID 11369855.

[14] Costantini S, Colonna G, Facchiano AM (2006). "Amino acid propensities for secondary structures are influenced by the protein structural class". *Biochem Biophys Res Commun.* 342 (2): 441–451. doi:10.1016/j.bbrc.2006.01.159. PMID 16487481.

Chou KC, Zhang CT (1995). "Prediction of protein structural classes". *Crit. Rev. Biochem. Mol. Biol.* 30 (4): 275–349. doi:10.3109/10409239509083488. PMID 7587280.

Guzzo, AV (1965). "Influence of Amino-Acid Sequence on Protein Structure". *Biophys. J.* 5 (6): 809–822. Bibcode:1965BpJ.....5..809G. doi:10.1016/S0006-3495(65)86753-4. PMC 1367904. PMID 5884309.

Prothero, JW (1966). "Correlation between Distribution of Amino Acids and Alpha Helices". *Biophys. J.* 6 (3):

367–370. Bibcode:1966BpJ.....6..367P. doi:10.1016/S0006-3495(66)86662-6. PMC 1367951. PMID 5962284.
Schiffer, M; Edmundson AB (1967). "Use of Helical Wheels to Represent Structures of Proteins and to Identify Segments with Helical Potential". *Biophys. J.* 7 (2): 121–35. Bibcode:1967BpJ.....7..121S. doi:10.1016/S0006-3495(67)86579-2. PMC 1368002. PMID 6048867.
Kotelchuck, D; Scheraga HA (1969). "The Influence of Short-Range Interactions on Protein Conformation, II. A Model for Predicting the α-Helical Regions of Proteins". *Proc Natl Acad Sci USA* 62 (1): 14–21. doi:10.1073/pnas.62.1.14. PMC 285948. PMID 5253650.
Lewis, PN; Gō N; Gō M; Kotelchuck D; Scheraga HA (1970). "Helix Probability Profiles of Denatured Proteins and Their Correlation with Native Structures"

[15] Froimowitz M, Fasman GD (1974). "Prediction of the secondary structure of proteins using the helix-coil transition theory". *Macromolecules* 7 (5): 583–9. doi:10.1021/ma60041a009. PMID 4371089.

[16] Dor O, Zhou Y (2006). "Achieving 80% tenfold cross-validated accuracy for secondary structure prediction by large-scale training". *Proteins* 66 (4): 838–45. doi:10.1002/prot.21298. PMID 17177203.

[17] Chou PY, Fasman GD (1974). "Prediction of protein conformation". *Biochemistry* 13 (2): 222–245. doi:10.1021/bi00699a002. PMID 4358940.

[18] Garnier J, Osguthorpe DJ, Robson B (1978). "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins". *J Mol Biol* 120 (1): 97–120. doi:10.1016/0022-2836(78)90297-8. PMID 642007.

[19] Robson B, Pain RH (May 1971). "Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions". *J. Mol. Biol.* 58 (1): 237–59. doi:10.1016/0022-2836(71)90243-9. PMID 5088928.

[20] Robson B (September 1974). "Analysis of code relating sequences to conformation in globular proteins. Theory and application of expected information". *Biochem. J.* 141 (3): 853–67. PMC 1168191. PMID 4463965.

[21] Pham TH, Satou K, Ho TB (2005). "Support vector machines for prediction and analysis of beta and gamma-turns in proteins". *J Bioinform Comput Biol* 3 (2): 343–358. doi:10.1142/S0219720005001089. PMID 15852509.

