African Journal of Biotechnology Vol. 7 (25), pp. 4687-4693, 29 December, 2008
Available online at http://www.academicjournals.org/AJB
ISSN 1684–5315 © 2008 Academic Journals

Review

Is protein structure prediction still an enigma?

K. Sobha[1]*, C. Kanakaraju[2] and K. Siva Krishna Yadav[2]

[1]Department of Biotechnology, RVR & JC College of Engineering, Chowdavaram, Guntur-522 019, Andhra Pradesh, India.
[2]Department of Biotechnology, Bapatla College of Engineering, Bapatla – 545 101, Andhra Pradesh, India.

Accepted 23 September, 2008

Proteins are large molecules indispensable for the existence and proper functioning of biological organisms. They perform a wide array of functions including catalysis, structure formation, transport, body defense, etc. Understanding the functions of proteins is a fundamental problem in the discovery of drugs to treat various diseases. The structure of a protein can be determined by physical methods which are slow and expensive but owing to the dramatic increase in the numbers of proteins sent to the public data bank during the last few years, it is highly desirable to develop some rapid and effective computational methods to predict the structure of new proteins so as to expedite the process of deducing their function. All the structure prediction methods basically rely on the idea that there is a correlation between residue sequence and structure. The primary structure is unique for each protein and it is generally accepted that a protein's primary structure is enough to determine its folding process to secondary, tertiary and quaternary structure. Despite recent efforts to develop automated protein structure determination protocols, structural genomic projects are slow in generating fold assignments for complete proteomes, and spatial structures remain unknown for many protein families. Alternative cheap and fast methods to assign folds using prediction algorithms continue to provide valuable structural information for many proteins. Protein structure determination and prediction has been a focal research subject in life sciences due to the importance of protein structure in understanding the biological and chemical activities of organisms/cell. This review comprehends the various recent advanced methods for protein structure predictions such as a two-stage method for assigning residues one of the three secondary structure states, prediction of homo-oligomeric proteins based on nearest neighbour algorithm, sequence–based hidden markov model, practical ab initio methods aimed at finding the native structure of the protein by simulating the biological process of protein folding, and metapredictors based on consensus form multiple methods.

Key words: Structure prediction, data classification, Hidden Markov model, homo-oligomeric proteins, nearest neighbour algorithm, twilight zone sequences.

INTRODUCTION

Proteins are polymer chains of repeating polypeptide units with side chains attached to each polypeptide unit. The side chains, also known as residues, are amino acids with different characteristics. There are 20 different amino acids in natural proteins. The sequence of amino acids in a protein chain is given by the primary structure. A typical protein contains 200 – 300 amino acids but this may increase up to approximately 30,000 in a single chain. The proteins have three local structural conformations: helices, sheets and other structural conformations such as loops, turns and coils. The secondary structure of proteins is the structural characterization of a protein with respect to these three local structural conformations.

The number of identified protein sequences has dramatically increased in the recent past due to the extensive research in the field. However, the majority of these sequences are not accompanied by any information about their function because of the slow and expensive nature of the experimental procedures for structure determination. So, one way to understand their function is to link them with known proteins in annotated databases, whose three-dimensional structure (fold) is

*Corresponding author. Email: sobha_kota@yahoo.co.in.

known. Even this structural knowledge is still limited because the experimental process of structure determination is slow for most proteins and not possible for many. The concern of the scientists at present is to abridge the gap between sequence and structure knowledge, often termed the sequence-structure gap. It is the main factor driving the need for predictions of protein structure. Many pharmaceutical drugs act by selective binding to target proteins and knowledge of protein structures can aid the process of rational structure based drug design. Levinthal (1966, 1968) believed that even for a small protein, there would be a large number of possible structures, and that the process of finding the correct one cannot proceed by a random search of the possibilities. But according to Anfinsen (1973), the proteins can fold to their native structures spontaneously without the intervention of any agent and therefore opined that protein fold is coded in the amino acid sequence itself. Protein structure prediction is therefore a problem of much scientific interest and it is still not clear as to how structure is encoded in sequence.

Computer methods for protein analysis address this problem since they study the relations within the amino acids sequence or structure. Since proteins have structural features which define functional similarities, the need for structure estimation methods is high. Protein computational analysis aims in structure estimation and includes two protein classification tasks: fold recognition and class prediction. Such methods are sequence-based or protein attribute-based. The availability of protein attributes' data is lower than the sequence information, either primary or secondary and so the focus is on exploiting sequence data for structure prediction. First predictions were made in 1970s with a few dozen structures available. Currently structures of about 51,491 (as of June 24, 2008) proteins are identified (stored in and accessible from Protein Data Bank Berman et al., 2000) providing good amount of data that supports more reliable predictions with better accuracy.

GENERAL PROTEIN STRUCTURE PREDICTION METHOD

The structure prediction methods can broadly be categorized as either *ab initio* or knowledge-based. Methods which explain protein folding based on calculation and minimization of free energy are called *ab initio* methods (Hardin et al., 2002). These methods use the accepted theories of quantum mechanics and statistical thermodynamics for prediction which in reality is just not possible because of the involvement of several thousands of atoms for which the free energy calculations are to be made. Knowledge-based methods attempt to predict structure of the unknown using the information from the database of known 3D structures. Comparative modeling, also called Homology modeling, is one of the many prediction methods and yields accurate results given the best alignments of the target (unknown) sequence and one or more template (known) sequences. Though different software packages perform the same homology modeling in slightly different ways, the basic steps are the same and include the following in order: (1) Analysis of template structure(s) and generation of an average if many templates are available (In the latter case, more similar sequences are more strongly weighted and finally a frame work of template atomic positions is calculated); (2) Generation of structure in two parts – first the core structure comprising the secondary structure elements and the second are the non-conserved loops; Structure prediction of non-conserved loops is difficult and is done with the aid of loop structure prediction algorithms (For details, the reader is advised to browse the internet for there are several sites). One such algorithm is the spare parts algorithm which makes use of a database of known loop structures from other proteins that may not necessarily be similar in sequence to the target. A spare part loop is selected from the database for each loop to be fitted into the gap in the modeled structure and is identified as the predicted loop. (These two steps complete the prediction of backbone atoms of the target sequence); (3) Side chains and their component atomic positions are predicted usually from a library of allowed side chain structures (Side chain rotamer library). With this step, the prediction of an overall model for the protein of interest is complete. But the predicted model can further be refined with the use of energy minimization soft wares without compromising the prediction accuracy of the model. Accuracy of a model is measured in terms of root mean square deviations between the α-carbon positions in the predicted structure and the actual structure of the target sequence. RMSDs of less than 1.0Å represent very good predictions and these values vary inversely with the % similarity in sequence. When it is not possible to get a full atom model of the tertiary structure, then the secondary structure predictions known as three-state predictions viz., helical, strand or extended or coil become useful.

Protein secondary structure prediction Methods

Two early methods for the prediction of secondary structure were those of Chou and Fasman (1974) and Garnier-Osguthorpe-Robson (GOR) (1978). These methods work on the basis of propensity of amino acids for attaining a particular structure. In the Chou - Fasman method, if in a run of six residues, four are helix favouring and the average value of the helix propensity is greater than 1.0 and greater than the average strand propensity, then the prediction is made as helix. Propensity values of amino acids for a particular secondary structure are calculated by dividing the frequency with which the particular residue is observed in the relevant secondary structure by the frequency for all residues in that secondary structure (These values are available in literature).

The secondary structure determination problem has been addressed using three categories of computational approaches: 1) comparative modeling; 2) threading; 3) *ab initio* methods (Baker and Sali, 2001). The computational methods that are based on comparative modeling exploit the fact that evolutionarily related proteins have similar sequences (Altschul et al., 1997). Threading compares a target sequence against a library of structural templates and produces a list of scores (Rost et al., 1997). The fold with the best score is assumed to be the one adopted by the sequence. The *ab initio* methods use only the sequence information of the protein to determine the protein structure. The objective in the *ab initio* methods is to determine the minimum free energy of the system. The free energy of the system depends on different interactions in the protein system such as ionic, non bonded, hydrogen-bonding and hydrophobic. The native conformation of the protein is the one corresponding to the coordinates of the atoms of the protein that gives the minimum free energy (Liwo et al., 1999; Bradley et al., 2003; Klepeis and Floudas 2003).

Among comparative modeling approaches to predict protein structures, the most successful ones include neural network models, database search tools, multiple sequence alignment, local sequence alignment, threading, Hidden Markov model-based methods, nearest neighbor methods, molecular dynamic simulation, and approaches combining different prediction methods. Neural networks are parallel, distributed information processing structures and the method tries to solve the problem by training the network (Bohr et al., 1990; Rost, 2001; Cai et al., 2002).

A host of Computational methods are developed to predict the location of secondary structure elements in proteins for complementing or creating insights into experimental results. However, prediction accuracies of these methods rarely exceed 70% (Rost and Sander, 1993).

Two stage method

The two-stage method presented here is a comparative method using the structure data available in the Protein Data Bank (Berman et al., 2000). Here a novel two-stage method to predict the location of secondary structure elements in a protein using the primary structure data only is presented. In the first stage of the proposed method, the folding type of a protein is determined using a novel classification approach for multi-class problems. The second stage of the method utilizes data available in the Protein Data Bank and determines the possible location of secondary structure elements in a probabilistic search algorithm. It is shown that the average accuracy of the predictions is 74.1% on a large structure data set.

Prediction of folding type

It was postulated by Nakashima et al. (1986) that the overall folding type of a protein depends on amino acid composition. Several methods are developed to exploit this postulate in the prediction of folding type of proteins (Chou, 1995; Bahar et al., 1997; Cai et al., 2001). These methods use statistical analysis and separate multi-dimensional amino acid composition data into several folding types. The prediction of protein folding type is a typical multi-class data classification problem. Classification of multi-dimensional data plays an important role in the decision to determine the main characteristics of a set. A support vector machine is a data mining method to classify data into different groups (Cai et al., 2001). Although this method can be efficient in classifying data into two groups, it is inaccurate and inefficient when the data needs to be classified into more than two sets. Mixed-integer programming allows the use of hyper boxes for defining boundaries of the sets that include all or some of the points in that set. Therefore, the efficiency and accuracy of multi-class data classification can be improved significantly compared to traditional methods (Turkay et al., 2005; Uney and Turkay, 2006).

Prediction of the secondary structure

Once the folding type of a protein is determined, the three state probabilities for each residue of the protein can be refined. For example, if a protein has an ‘all-α” folding type, then it is not possible for any residue of this protein to be in a “β-sheet” conformation. In order to exploit this fact in the two-stage method, the protein data set is divided into four subsets: “all-α”, “all-β”, “α/β” and α+β”. The folding type of each protein is obtained from the SCOP (Murzin et al., 1995) database and these four separate databases are used to calculate the frequency of occurrences of each amino acid in a α-helix, β- sheet or other structures. The basis for the algorithm is searching segments of its residue sequence in pool of known protein structures and predicting structure for each resi-due on the basis of frequency of occurrence. To determine the number of residues in each segment to be searched, two factors are considered: the segment should be long enough to have a legitimate reason to search considering interactions and the bonds formed between amino acids to shape their structures. The partitioning of the protein chain into overlapping segments of oligo peptide chains was suggested by Anfinsen and Scheraga (1975). These authors used segments of nine residues in their free energy calculations. The use of overlapping segments with five residues is shown to be very effective in predicting the helical segments of proteins

An important consideration in establishing databases for calculation of three state probabilities is the redundant data. Some proteins in organisms are coded by the same gene and are very similar in terms of residue sequence, structure, and function. These molecules are called “homologous” and can be considered to contain the same information due to the very high level of similarity mentioned. While searching a segment of residue sequence of a protein in the database, occurrence of segment in homologous structures can bias the results. That is, although homologous structures contain the same information, all occurrences of segment in the database are being counted, which is similar to utilizing the same data more than once. When the probability is calculated for a particular segment being part of the structure that is the same with that of homologous for the residues in the segment, the probabilities are biased for homologous sequences. Therefore, it is necessary to exclude homologous structures from databases, which are carried out by queries in the database. Then, the probability for a particular residue to be in a secondary element for each residue segment is calculated according to its folding type. Secondary structure of the peptide hormone called Eclosion Hormone (EH) (a neuropeptide of 62 amino acid residues in insects controlling ecdysis) was predicted using the average distance map method (ADM) and was shown to contain an N-terminal helix which is active in the construction of globular structure of the peptide. Subsequently, C-terminal structure was predicted by a method complementary to ADM and the functional residues were analysed using glycine-substitution technique. Finally, the 3D structure of the peptide was constructed by computer aided modeling and energy minimization using Insight-II/Discover software (Fujita et al., 1998).

Sequence-based protein structure prediction using a reduced state-space hidden Markov model

This method (Lampros et al., 2007) describes the use of a hidden Markov model (HMM), with a reduced number of states, which simultaneously learns amino acid sequence and secondary structure for proteins of known three-dimensional structure and performs two tasks: protein class prediction and fold recognition. The protein data bank and the annotation of the SCOP database are used for training and evaluation of the proposed HMM for a number of protein classes and folds. Results demonstrate that the reduced state–space HMM performs equivalently or even better in some cases, on classifying proteins than a HMM trained with the amino acid sequence. The major advantage of the proposed approach is that a small number of states are employed and the training algorithm is of low complexity and thus relatively fast.

A review of sequence-based approaches reveals that hidden Markov models (HMMs) are those most commonly used and also demonstrate high performance. HMMs have been applied for multi-class protein fold recognition (Lindahl and Elofsson, 2000) employing the sequence alignment and modeling (SAM) software (Hughey and Krogh, 1996). Furthermore, secondary structure information can be incorporated in the HMM and increase the fold recognition performance (Hargbo and Elofsson, 1999). Karchin et al. (2003) used the same approach and additionally they have evaluated different alphabets for backbone geometry and their effect on the classification performance. However, the main drawback of HMMs is the employment of large model architectures which require large data sets and high computational effort for training. As a consequence, in cases where these data sets are not available, e.g. small classes or folds, their performance deteriorates. Very recently, the gene encoding eclosion hormone of the asian corn borer, *Ostrinia furnacalis* was sequenced and its molecular characteristics along with expression analysis were elucidated (Wei et al., 2008). In this study, the 3D structure of Osf-EH was modeled using HMMSTR prediction server and the hormone which is a 62 amino acid mature peptide is predicted to have four β-turns and three α-helices with the pattern of 2β-2α-2β-α.

PREDICTION OF PROTEIN STRUCTURAL CLASS FOR THE TWILIGHT ZONE SEQUENCES

The challenging problem of structural class prediction for the twilight zone sequences is addressed by a novel approach that aims to improve the prediction accuracy via designing a composite sequence representation that includes amino acid composition, physico-chemical properties and predicted secondary structure content (Kurgan and Chen, 2007). Their proposed LLSC-PRED method applies easy to comprehend and fast to train linear logistic regression classifier in comparison with the support vector machine based classifier. The sequence representation including 58 features together with the transparent prediction model aid in exposing subtle relationships between important physico-chemical sequence properties and the structural classes.

This method is an accurate method for in-silico prediction of structural classes from low homology (twilight zone) protein sequences. The proteins characterized by a lower, 20 – 30%, homology with sequences that are used to predict their structure are called the twilight zone proteins (Rost, 1999). More than 95% of all sequence pairs detected in the twilight zone have different structures (Rost, 1999), which significantly reduces the prediction accuracy. For instance, prediction of the secondary structure for homologous sequences by the state-of-the-art alignment-based methods yields about 80% accuracy, while for the twilight zone sequences it drops to only 65 – 68% (Lin et al., 2005) Similarly in case of structural class prediction, accuracies for highly homologous protein datasets reach over 90%, while they drop to about 57% in case of the twilight zone sequences (Kurgan and Homacian, 2006).

Structural class categorizes various proteins into groups that share similarities in the local folding. Prediction of structural classes is based on identifying these folding patterns based on thousands of already categorized proteins, and applying these patterns to millions of proteins with unknown structures but known amino acid (AA) sequences. (July 11, 2008 release 30 of the NCBI's RefSeq database stores 5, 590,364 amino acid (protein) sequences from 5,395 organisms). One of the most accurate classifications of structural classes can be found in the expert-curated SCOP (Structural Classification of Proteins) database (Murzin et al., 1995) (as of 26, September 2007, release 1.73 of SCOP stores 34,494 PDB entries; 97, 178 domains; Total no. of folds: 1086; No. of super families: 1777; No. of families: 3464). The basic structural unit of classification in SCOP database is either the entire sequence or a protein domain (structurally conserved fragment of the sequence). The database is organized as a hierarchy of known protein and protein domain structures where first level is based on the structural class: all-α, all-β, α/β and $\alpha+\beta$. Structural class prediction is usually performed in two steps - transformation of AA sequences into a fixed-length feature vectors and then the feeding of feature vectors to a classification algorithm to perform the prediction. Majority of the recent methods developed for prediction of structural classes include algorithms such as fuzzy clustering, neural net works, logistic regression, decision tree, Support vector machine (SVM) and the most recent complex classification models such as ensembles, bagging, boosting etc. All these different methods require atleast ~30% homology between the query sequence and the template and are not suitable for low homology sequences. For the twilight zone proteins, the prediction accuracies with the above algorithms for secondary structure and structural classes drops significantly as mentioned above.

PREDICTION OF HOMO-OLIGOMERIC PROTEINS BASED ON NEAREST NEIGHBOUR ALGORITHM

The primary structure which is unique for each protein is believed to determine its own fold and combination with other proteins to make the appropriate secondary, tertiary and quaternary structure (Anfinsen et al., 1961; Anfinsen, 1973). Proteins with quaternary protein structure are said to be oligomeric (or multimeric), and the individual chains are called subunits. Oligomeric proteins are either homo-oligomeric, consisting of identical subunits or hetero-oligomeric, consisting of different subunits. Arrangement of subunits in the oligomeric structure can also vary. An oligomeric protein is more than the sum of its parts and have important properties not shared with its separated subunits. A variety of bonding interactions including hydrogen bonding, salt bridges, and disulfide bonds hold the various subunits into a particular geometry. Klotz et al. (1975) reviewed a number of quaternary structure properties such as stoichiometric constitution, the geometric arrangements of the subunits, the assembly energetics, inter subunit communication, and their functional aspects. Some recent works have paid more attention to analyzing protein–protein interactions and predicting interactions sites (Marcotte et al., 1999; Bock and Gough, 2001; Glaser et al., 2001; Nooren and Thornton, 2003; Ofran and Rost, 2003). Results of the study by Garian (2001) using decision tree model and amino acid indices discriminating homo and non-homodimers at the level of primary sequences itself confirmed that the protein primary structure contains information on the quaternary structure of the same. Further studies by Zhang et al. (2003) using the support vector machine (SVM) and the covariant discriminant algorithms revealed that SVM is superior to both covariant discriminant algorithm and decision tree method as well. Nearest neighbour algorithm (Friedman et al., 1975) with subsequence distribution (Fang, 1994; Fang et al., 2001) is applied to discriminate homo-oligomeric proteins from the protein primary structure (Song, 2007). This algorithm describes protein primary sequences by their subsequence distributions. When the length of subsequence increases to appropriate level (l = 3,4), all the four performance measures viz., overall accuracy (Q), true positive rate (TPR), false positive rate (FPR) and Matthews correlation coefficient (MCC) were found to ascend quickly for the tests of the method on the chosen data set (a subset R_{1568} of Robert Garian's R_{1639}). The tests demonstrated that the residue order along protein sequences plays an important role in recognition of the homo-oligomers, and nearest neighbour algorithm method is a simple and effective tool for classification of homo-oligomeric proteins. It is further confirmed that protein primary sequence encodes quaternary structure information.

The nearest neighbour algorithm tries to classify the new patterns into their class membership by comparing the features of the unknown new patterns with the features of the patterns which have already been classified (Friedman et al., 1975). It is particularly useful in the situations when the distributions of the patterns and the categories of the patterns are unknown. The method will weigh heavily the evidence derived from the nearby patterns and is attractive because of its simplicity and low probability of error.

COMPLEX-TYPE-DEPENDENT SCORING FUNCTIONS IN PROTEIN–PROTEIN DOCKING

Protein-protein interactions, the basis of many biological regulations, are dependent on their 3D structures. In the case of large macromolecular assemblies, the amount of experimental structures of Protein-protein complexes is relatively quite small and cost expensive and therefore a combination of protein modeling and experimental structure determination increases knowledge of structure-based analysis of the protein-protein interaction network (Tovchigreehko et al., 2002; Chance et al., 2002; Sali et al., 2003; Janin and Levitt, 2006). A major challenge in the field of protein–protein docking is to discriminate between the many wrong and few near-native conformations, i.e. scoring. Wide usage of docking algorithms depends on their ability to generate potential structures and a good scoring function to distinguish the near-native structures from a large number of non-native ones. The presently used scoring functions surface include complementarity (Katchalski-Katzir et al., 1992; Walls and Sternberg, 1992; Ma et al., 2005), surface complementarity together with an electrostatic filter (Gabb et al., 1997; Heifetz et al., 2002), knowledge- based statistical potential such as atomic contact energy (ACE) (Zhang et al., 1997), the residue pair potential (RP) (Moont et al., 1999) and DFIRE (Liu et al., 2004). Some combinatorial functions are used in docking prediction (Fernandez-Recio et al., 2003; Gray et al., 2003) but none proved to be robust enough to all types of protein-protein complexes. The complex type-dependent combinatorial scoring functions developed for protease/inhibitor, enzyme/ inhibitor, antibody/antigen and other complexes incorporate both physical and knowledge-based potentials. The weights of the scoring functions for different type complexes were optimized by the multiple linear regression method (Li et al., 2007). The training set was constructed comprising only top 300 structures with ligand root mean square deviation less than 20Å. The study included bound-docking to examine the quality of the scoring function and was also extended to unbound-docking studies.

The physico-chemical characteristics of the binding interface vary with different types of protein complexes and studies of Li et al. (2007) considered the biological function of the complex as the principal factor. Protease/ inhibitors were segregated from enzyme/ inhibitors because of distinctive differences between the two complexes. The protease/inhibitor complex has a more hydrophobic interface than the enzyme/inhibitor complex. The scoring functions were effectively optimized in order to distinguish the near-native structures from non-native ones for different types of protein-protein complexes. Li et al. (2007) divided the protein-protein complexes into 4 categories and designed the scoring function with the regression method for each type and got 4 combinatorial functions that exhibited certain aptitude to select hit structures from all docked solutions. Results of this study take the scoring function development a step forward as the used scoring functions show relatively good abilities in distinguishing hit structures.

CONCLUSIONS

It is obvious that determining the structure of all identified proteins and deducing their function from the structures by physical means is extremely difficult as protein crystallization is not an easy task. In this context, the field of research encompassing development of computational approaches for determining structures based on amino acid sequence and comparing with appropriate templates of the protein data bank has come into existence and is growing fast with several approaches developed and in recent past, many of the protein structures have been deposited in PDB based on computational methods. But the key features that structure does not always complement function, and the microenvironment of the protein determines the folding pattern and ultimately its 3D structure should not be ignored. Simple biochemical and genetical methods are to be developed to confirm and ensure that the predicted protein structure by algorithms/bioinformatic tools is accurate enough in deducing the function of the protein under consideration. As and when a new protein is isolated and its sequence determined, then the structure of the protein be elucidated by computational approaches considering the cell environment from which the protein has been isolated and deduce the function by genetical and biochemical studies.

REFERENCES

Altschul S, Madden T, Shaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997). Gapped Blast and PSI-Blast: A new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Anfinsen CB (1973). Principles that govern the folding of protein chains. Science, 181: 223-230.

Anfinsen CB, Scheraga HA (1975). Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, 29: 205-300.

Anfinsen CB, Haber E, Scla M, White EH (1961).The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Nat. Acad. Sci. USA, 47:1309-1314.

Bahar I, Atilgan AR, Jernigan RL, Erman B (1997). Understanding the Recognition of Protein Structural Classes by Amino Acid Composition. Proteins: Struct. Funct. Genet. 29: 172-185.

Baker D, Sali A (2001). Protein structure prediction and structural genomics. Science, 294(5540): 93-96.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank. Nucleic Acids Res., 28(1): 235-242. (Protein Data Bank, htttp://www.pdb.org/).

Bock JR, Gough DA (2001). Predicting protein–protein interactions from primary structure. Bioinformatics 17: 455–460.

Bohr H, Bohr J, Brunak S, Cotterill RMJ, Fredholm H, Lautrup B Petersen SB (1990). A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. FEBS Lett., 261: 43-46.

Bradley P, Chivian D, Meiler J, Misura KM, Rohl C, Schief W, Wedemeyer WJ, Schueler-FO, Murphy P, Schonbrun J, Strauss C, Baker D (2003): Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation, Proteins: Struct. Funct. Genet. 53: 457-68.

Cai YD, Liu XJ, Xu XB Chou KC (2002). Artificial neural network method for predicting protein secondary structure content. Comput. Chem. 26: 347-350.

Cai YD, Liu XJ, Xu XB, Zhou GP(2001). Support Vector Machines for predicting protein structural class. BMC Bioinformatics, pp. 2, 3.

Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK (2002). Structural genomics: a pipeline for providing structures for the biologist. Protein Sci., 11: 723-738.

Chou KC (1995). Does the folding type of a protein depend on its amino acid composition? FEBS Lett., 363: 127-131.

Chou PY, Fasman GD (1974). Prediction of protein conformation. Biochemistry. 13(2):222-45.

Fang W (1994). The disagreement degree of multi-person judgements in additive structure. Math. Soc. Sci., 28: 85-111.

Fang W, Roberts ES, Ma Z (2001). A measure of discrepancy of multiple sequences. Inf. Sci. 137: 75-102.

Fernandez-RJ, Totrov M, Abagyan R (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. Proteins: Struct. Funct. Biol. 52: 113-117.

Friedman JH, Baskett F, Shustek LJ (1975). An algorithm for finding nearest neighbors. IEEE Trans. Comput., 24: 1000-1006.

Fujita N, Maekawa T, Ohta S, Kikuchi T (1998). The Functional residues and their representation by a hypothetical 3D model of silk worm eclosion hormone. Protein Eng. 11(9): 769-773.

Gabb HA, Jackson RM, Sternberg MJ (1997). Modeling protein docking using shape complementarity, electrostatics and biochemical information. J. Mol. Biol. 272: 06-120.

Garian R (2001). Prediction of quaternary structure from primary structure. Bioinformatics, 17: 551-556.

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: 97-120.

Glaser F, Steinberg DM, Vakser IA, Ben-TN (2001). Residue frequencies and pairing preference at protein–protein interfaces. Proteins Struct. Funct. Genet. 43: 89-102.

Gray JJ, Moughan SE, Wang C, Schueler-FO, Kuhlman B, Rohl CA, Baker D (2003). Protein–protein docking with simultaneous optimization of rigid body displacement and side chain conformations. J. Mol. Biol. 331: 281-299.

Hardin C, Pogorelov TV, Schulten ZL (2002). *Ab initio* protein structure prediction. Curr. Opin. Struct. Biol. 12(2): 176-181.

Hargbo J, Elofsson A (1999). Hidden Markov models that use predicted secondary structures for fold recognition. Proteins, 36: 68-76.

Heifetz A, Katchalski-KE, Eisenstein M (2002). Electrostatics in protein–protein docking. Protein Sci. 11: 571-587.

Hughey R, Krogh A (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. CABIOS 12 (2): 95-107.

Janin J, Levitt M (2006). Theory and simulation accuracy and reliability in modeling proteins and complexes. Curr. Opin. Struct. Biol. 16: 1-3.

Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K (2003). Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins, 51: 504-514.

Katchalski-Katzir E, Shariv I, Eisenstein MA, Friesem, Aflalo C, Vakser I (1992). Molecular surface recognition: determination of geometric fit between protein and their ligands by correlation techniques. Proc. Natl. Acad. Sci. USA, 89: 2195-2199.

Klepeis JL, Floudas CA (2003). ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. Biophys. J., 85: 2119-2146.

Klotz IM, Darnall DM, Langerman NR, Quaternary structure of proteins, In: *The Proteins*, Vol.1 H. Neurath RL Hill (Eds.), Academic Press, New York, 1975, pp.293-411.

Kurgan L, Homacian L (2006). Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recogn. 39 (12): 2323-2343.

Kurgan L, Chen K (2007). Prediction of protein structural class for the twilight zone sequences. Biochem. Biophys. Res. Commun. 357: 453-460.

Lampros C, Papaloukas C, Exarchos TP, Goletsis Y, Fotiadis DI (2007). Sequence-based protein structure prediction using a reduced state-space hidden Markov model. Comput. Biol. Med. 37: 1211-1224.

Levinthal C (1966). Molecular model-building by computer. Sci. Am. 214: 42-52.

Levinthal C (1968). Are there pathways for protein folding? J. Chem. Phys. 65: 44-45.

Li CH, Ma XH, Shen LZ, Chang S, Chen WZ, Wang CX (2007). Complex-type- dependent scoring functions in protein-protein docking. Biophys. Chem. 129: 1-10.

Lin K, Simossis VA, Taylor WR, Heringa J (2005). A simple and fast secondary structure rediction method using hidden neural networks. Bioinformatics, 21(2): 152-159.

Lindahl E, Elofsson A (2000). Identification of related proteins on family, super family and fold level. J. Mol. Biol., 295: 613-625.

Liu S, Zhang C, Zhou H, Zhou Y (2004). A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins: Struct. Funct. Biol. 56: 93-101.

Liwo A, Lee J, Ripoll D, Pillardy J, Scheraga H (1999). Protein structure prediction by global Optimization of a potential energy function. Proc. Natl. Acad. Sci. USA, 96(10): 5482-5485.

Ma XH, Li CH, Shen LZ, Gong XQ, Chen WZ, Wang CX (2005). Biologically enhanced sampling geometric docking and backbone flexibility treatment with multiconformational superposition. Proteins: Struct. Funct. Biol. 60: 319-323.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999). Detecting protein function and protein–protein interactions from genome sequences. Science, 285: 751-753.

Moont G, Gabb HA, Sternberg MJE (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins: Struct. Funct. Biol. 35: 364-373.

Murzin A, Brenner S, Hubbard T, Chothia C (1995). SCOP: a structural classification of protein database for the investigation of sequence and structures. J. Mol. Biol., 247: 536-540.

Nakashima H, Nishikawa K, Ooi T (1986). The folding type of a protein is relevant to the amino-acid composition. J. Biochem., 99: 152-162.

Nooren IMA, Thornton JM ((2003). Structural characterization and functional significance of transient protein–protein interactions. *J. Mol. Biol*. 325: 991–1018.

Ofran Y, Rost B (2003). Analyzing six types of protein–protein interfaces. J. Mol. Biol. 325: 377-387.

Rost B (1999). Twilight zone of protein sequence alignments. Protein Eng. 2: 85-94.

Rost B (2001) Review: Protein secondary structure prediction continues to rise. J. Struct. Biol., 134: 204-218.

Rost B, Sander C (1993). Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol., 232: 584-599.

Rost B, Schneider R, Sander C (1997). Protein fold recognition by prediction based threading. J. Mol. Biol., 270(3): 471-480.

Sali A, Glaeser R, Earnest T, Baumeister W (2003). From words to literature in structural proteomics. Nature, 422: 216-225.

Song J (2007). Prediction of homo-oligomeric proteins based on nearest neighbour algorithm. Comput. Biol. Med, 37(12): 1759-1764.

Tovchigreehko A, Well CA, Vakser IA (2002). Docking of Protein models. Protein Sci. 11: 1888-1896.

Turkay M, Uney F, Yilmaz O (2005). Prediction of Folding Type of Proteins Using Mixed-Integer Linear Programming, In Computer Aided Chem. Eng., vol. 20A: ESCAPE-15, L Puigjaner and A Espuna (Eds.), Elsevier Publishers, pp. 523-528.

Uney F, Turkay M (2006). A Mixed-Integer Programming Approach to Multi-Class Data Classification Problem. Eur. J. Oper. Res., 173(3): 910-920.

Walls PH, Sternberg MJ (1992). New algorithm to model protein–protein recognition based on surface complementarity: applications to antibody – antigen docking. J. Mol. Biol. 228: 277-297.

Wei ZJ, Hong GY, Wei HY, Jiang ST, Lu C (2008). Molecular characters and expression analysis of the gene encoding eclosion hormone from the Asian corn borer, *Ostrinia furnacalis.* DNA Sequence, 19(3): 301-307.

Zhang C, Vasmatzis G, Cornette JL, Delisi C (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. J. Mol. Biol. 267: 707– 726.

Zhang S, Pan Q, Zhang H, Zhang Y and Wang H (2003). Classification of protein quaternary structure with support vector machine. Bioinformatics, 18: 2390-2396.