CHAPTER 1

INTRODUCTION TO PROTEIN STRUCTURE PREDICTION

HUZEFA RANGWALA
Department of Computer Science
George Mason University
Fairfax, VA

GEORGE KARYPIS
Department of Computer Science
University of Minnesota
Minneapolis, MN

Proteins have a vast influence on the molecular machinery of life. Stunningly complex networks of proteins perform innumerable functions in every living cell. Knowing the function and structure of proteins is crucial for the development of improved drugs, better crops, and even synthetic biofuels. As such, knowledge of protein structure and function leads to crucial advances in life sciences and biology.

With recent advances in large-scale sequencing technologies, we have seen an exponential growth in protein sequence information. Protein structures are primarily determined using X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, but these methods are time consuming, expensive, and not feasible for all proteins. The experimental approaches to determine protein function (e.g., gene knockout, targeted mutation, and inhibitions of gene expression studies) are low-throughput in nature [1,2]. As such, our ability to produce sequence information far outpaces the rate at which we can produce structural and functional information.

Consequently, researchers are increasingly reliant on computational approaches to extract useful information from experimentally determined three-dimensional (3D) structures and functions of proteins. Unraveling the

Introduction to Protein Structure Prediction: Methods and Algorithms,
Edited by Huzefa Rangwala and George Karypis

relationship between pure sequence information and 3D structure and/or function remains one of the fundamental challenges in molecular biology.

Function prediction is generally approached by using inheritance through homology [2], that is, proteins with similar sequences (common evolutionary ancestry) frequently carry out similar functions. However, several studies [2–4] have shown that a stronger correlation exists between structure conservation and function, that is, structure implies function, and a higher correlation exists between sequence conservation and structure, that is, sequence implies structure (sequence → structure → function).

1.1. INTRODUCTION TO PROTEIN STRUCTURES

In this section we introduce the basic definitions and facts about protein structure, the four different levels of protein structure, as well as provide details about protein structure databases.

1.1.1. Protein Structure Levels

Within each structural entity called a protein lies a set of recurring substructures, and within these substructures are smaller substructures still. As an example, consider hemoglobin, the oxygen-carrying molecule in human blood. Hemoglobin has four domains that come together to form its quaternary structure. Each domain assembles (i.e., folds) itself independently to form a tertiary structure. These tertiary structures are comprised of multiple secondary structure elements—in hemoglobin's case α-helices. α-Helices (and their counterpart β-sheets) have elegant repeating patterns dependent upon sequences of amino acids.

1.1.1.1. Primary Structure. Amino acids form the basic building blocks of proteins. Amino acids consists of a central carbon atom (C_α) attached by an amino (NH_2), a carboxyl (COOH) group, and a side chain (R) group. The side chain group differentiates the various amino acids. In case of proteins, there are primarily 20 different amino acids that form the building blocks. A protein is a chain of amino acids linked with peptide bonds. Pairs of amino acid form a peptide bond between the amino group of one and the carboxyl group of the other. This polypeptide chain of amino acids is known as the primary structure or the protein sequence.

1.1.1.2. Secondary Structure. A sequence of characters representing the secondary structure of a protein describes the general 3D form of local regions. These regions organize themselves independently from the rest of the protein into patterns of repeatedly occurring structural fragments. The most dominant local conformations of polypeptide chains are α-helices and β-sheets. These local structures have a certain regularity in their form, attributed to the hydrogen bond interactions between various residues. An α-helix has a coil-like

structure, whereas a β-sheet consists of parallel strands of residues. In addition to regular secondary structure elements, irregular shapes form an important part of the structure and function of proteins. These elements are typically termed coil regions.

Secondary structure can be divided into several types, although usually at least three classes (α-helix, coils, and β-sheet) are used. No unique method of assigning residues to a particular secondary structure state from atomic coordinates exists, although the most widely accepted protocol is based on the Dictionary of Protein Secondary Structure (DSSP) algorithm [5]. DSSP uses the following structural classes: H (α-helix), G (3_{10}-helix), I (π-helix), E (β-strand), B (isolated β-bridge), T (turn), S (bend), and – (other). Several other secondary structure assignment algorithms use a reduction scheme that converts this eight-state assignment down to three states by assigning H and G to the helix state (H), E and B to a the strand state (E), and the rest (I, T, S, and –) to a coil state (C). This is the format generally used in structure databases.

1.1.1.3. Tertiary Structure. The tertiary structure of the protein is defined as the global 3D structure, represented by 3D coordinates for each atoms. These tertiary structures are comprised of multiple secondary structure elements, and the 3D structure is a function of the interacting side chains between the different amino acids. Hence, the linear ordering of amino acids forms secondary structure; arranging secondary structures yields tertiary structure.

1.1.1.4. Quaternary Structure. Quaternary structures represent the interaction between multiple polypeptide chains. The interaction between the various chains is due to the non-covalent interactions between the atoms of the different chains. Examples of these interactions include hydrogen bonding, van Der Walls interactions, ionic bonding, and disulfide bonding.

Research in computational structure prediction concerns itself mainly with predicting secondary and tertiary structures from known experimentally determined primary structure or sequence. This is due to the relative ease of determining primary structure and the complexity involved in quaternary structure.

1.1.2. Protein Sequence and Structure Databases

The large amount of protein sequence information, experimentally determined structure information, and structural classification information is stored in publicly available databases. In this section we review some of the databases that are used in this field, and provide their availability information in Table 1.1.

1.1.2.1. Sequence Databases. The Universal Protein Resource (UniProt) [6] is the most comprehensive warehouse containing information about protein

TABLE 1.1 Protein Sequence and Structure Databases

Database	Information	Availability Link
UniProt	Sequence	http://www.pir.uniprot.org/
UniRef	Cluster sequences	http://www.pir.uniprot.org/
NCBI nr	Nonredundant sequences	ftp://ftp.ncbi.nlm.nih.gov/blast/db/
PDB	Structure	http://www.rcsb.org/
SCOP	Structure classification	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH	Structure classification	http://www.cathdb.info/
FSSP	Structure classification	http://www.ebi.ac.uk/dali/fssp/
ASTRAL	Compendium	http://astral.berkeley.edu/

The databases referred to in this table are most popular for protein structure-related information.

sequences and their annotation. It is a database of protein sequences and their function that is formed by aggregating the information present in the Swiss-Prot, TrEMBL, and Protein Information Resources (PIR) databases. The UniProtKB 13.2 version of database (released on April 8, 2008) consists of 5,939,836 protein sequence entries (Swiss-Prot providing 362,782 entries and TrEMBL providing 5,577,054 entries).

However, several proteins have high pairwise sequence identity, and as such lead to redundant information. The UniProt database [6] creates a subset of sequences such that the sequence identity between all pairs of sequences within the subset is less than a predetermined threshold. In essence, UniProt contains the UniRef100, UniRef90, and UniRef50 subsets where within each group the sequence identity between a pair of sequences is less than 100%, 90%, and 50%, respectively.

The National Center for Biotechnology Information (NCBI) also provides a nonredundant (NCBI nr) database of protein sequences using sequences from a wide variety of sources. This database will have pairs of proteins with high sequence identity, but removes all the duplicates. The NCBI nr version 2.2.18 (released on March 2, 2008) contains 6,441,864 protein sequences.

1.1.2.2. Protein Data Bank (PDB). The Research Collaboratory for Structural Bioinformatics (RSCB) PDB [7] stores experimentally determined 3D structure of biological macromolecules including nucleotides and proteins. As of April 20, 2008 this database consists of 46,287 protein structures that are determined using X-ray crystallography (90%), NMR (9%), and other methods like Cryo-electron microscopy (Cryo-EM). These experimental methods are time-consuming, expensive, and need protein to crystallize.

1.1.2.3. Structure Classification Databases. Various methods have been proposed to categorize protein structures. These methods are based on the pairwise structural similarity between the protein structures, as well as the topological and geometric arrangement of atoms and predominant secondary

structure like subunits. Structural Classification of Proteins (SCOP) [8], Class, Architecture, Topology, and Homologous superfamily (CATH) [9], and Families of Structurally Similar Proteins (FSSP) [10] are three widely used structure classification databases. The classification methodology involves breaking a protein chain or complex into independent folding units called domains, and then classifying these domains into a set of hierarchical classes sharing similar structural characteristics.

SCOP Database. SCOP [8] is a manually curated database that provides a detailed and comprehensive description of the evolutionary and structural relationships between proteins whose structure is known (present in the PDB). SCOP classifies proteins structures using visual inspection as well as structural comparison using a suite of automated tools. The basic unit of classification is generally a domain. SCOP classification is based on four hierarchical levels that encompass evolutionary and structural relationships [8]. In particular, proteins with clear evolutionary relationship are classified to be within the same *family*. Generally, protein pairs within the same family have pairwise residue identities greater than 30%. Protein pairs with low sequence identity, but whose structural and functional features imply probably common evolutionary information, are classified to be within the same *superfamily*. Protein pairs with similar major secondary structure elements and topological arrangement of substructures (as well as favoring certain packing geometries) are classified to be within the same *fold*. Finally, protein pairs having a predominant set of secondary structures (e.g., all α-helices proteins) lie within the same *class*. The four hierarchical levels, that is, family, superfamily, fold, and class define the structure of the SCOP database.

The SCOP 1.73 version database (released on September 26, 2007) classifies 34,494 PDB entries (97,178 domains) into 1086 unique folds, 1777 unique superfamilies, and 3464 unique families.

CATH Database. CATH [9] database is a semi-automated protein structure classification database like the SCOP database. CATH uses a consensus of three automated classification techniques to break a chain into domains and classify them in the various structural categories [11]. Domains for proteins that are not resolved by the consensus approach are determined manually. These domains are then classified into the following hierarchical categories using both manual and automated methods in conjunction.

The first level membership, *class*, is determined based on the secondary structure composition and packing within the structure. The second level, *architecture*, clusters proteins sharing the same orientation of the secondary structure element but ignoring the connectivity between these substructural units. The third level, *topology*, groups protein pairs with a high structure alignment score as determined by the SSAP [12] algorithm, and in essence share both overall shape and connectivity of secondary structures. The fourth level, *homologous* pairs, shares a common ancestor and is identified by

sequence alignment as well as the SSAP structure alignment method. Structures are further classified to be within the same *sequence families* if they share a high sequence identity.

The CATH 3.1.0 version database (released on January 19, 2007) classifies 30,028 (93,885 domains) proteins from the PDB into 40 architecture-level classes, 1084 topology-level classes, and 2091 homologous-level classes.

FSSP Database. The FSSP [10] is a structure classification database. FSSP uses an automatic classification scheme that employs exhaustive structure-to-structure alignment of proteins using the DALI [13] alignment. FSSP does not provide a hierarchical classification like the SCOP and CATH databases, but instead employs a hierarchical clustering algorithm using the pairwise structure similarity scores that can be used for the definition of fold classes—however, not very accurate.

There have been several studies [14,15] analyzing the relationship between the SCOP, CATH, and FSSP databases for representing the fold space for proteins. The major disagreement between the three databases lies in the domain identification step, rather than the domain classification step. A high percentage of agreement exists between the SCOP, CATH, and FSSP databases especially at the fold level with sequence identity greater than 25%.

ASTRAL Compendium. The A Structural Alignment Library (ASTRAL) [16–18] compendium is a set of database and tools used for analysis of protein structures and sequences. This database is partially derived from, and augments, the SCOP [8] database. ASTRAL provides accurate linkage between the biological sequence and the reported structure in PDB, and identifies the domains within the sequence using SCOP. Since the majority of domain sequences in PDB are very similar to others, ASTRAL tools reduce the redundancy by selecting high-quality representatives. Using the reduced nonredundant set of representation proteins allows for sampling of all the different structures in the PDB. This also removes bias due to overrepresented structures. Subsets provided by ASTRAL are based on SCOP domains and use high-quality structure files only. Independent subsets of representative proteins are identified using a greedy algorithm with filtering criterion based on pairwise sequence identity determined using the Basic Local Alignment Search Tool (BLAST) [19], an e-value-based threshold, or a SCOP level-based filter.

1.2. PROTEIN STRUCTURE PREDICTION METHODS

One of the biggest goals in structural bioinformatics is the prediction of the 3D structure of a protein from its one-dimensional (1D) protein sequence. The goal is to be able to determine the shape (known as a fold) that a given amino acid sequence will adopt. The problem is further divided based on

whether the sequence will adopt a new fold or bear resemblance to an existing fold (template) in some protein structure database. Fold recognition is easy when the sequence in question has a high degree of sequence similarity to a sequence with known structure [20]. If the two sequences share evolutionary ancestry they are said to be homologous. For such sequence pairs we can build the structure for the query protein by choosing the structure of the known homologous sequence as template. This is known as comparative modeling.

In the case where no good template structure exists for the query, one must attempt to build the protein tertiary structure from scratch. These methods are usually called *ab initio* methods. In a third-fold prediction scenario, there may not necessarily be a good sequence similarity with a known structure, but a structural template may still exist for the given sequence. To clarify this case, if one were aware of the target structure then they could extract the template using structure–structure alignments of the target against the entire structural database. It is important to note that the target and template need not be homologous. These two cases define the fold prediction (homologous) and fold prediction (analogous) problems during the Critical Assessment of Protein Structure Prediction (CASP) competition.

1.2.1. Comparative Modeling

Comparative Modeling or homology modeling is used when there exists a clear relationship between the sequence of a query protein (unknown structure) and a sequence of a known structure. The most basic approach to structure prediction for such (query) proteins is to perform a pairwise sequence alignment against each sequence in protein sequence databases. This can be accomplished using sequence alignment algorithms such as Smith-Waterman [21] or sequence search algorithms (e.g., BLAST [19]). With a good sequence alignment in hand, the challenge in comparative modeling becomes how to best build a 3D protein structure for a query protein using the template structure.

The heart of the above process is the selection of a suitable structural template based on sequence pair similarity. This is followed by the alignment of query sequence to the template structure selected to build the backbone of the query protein. Finally the entire modeled structure is refined by loop construction and side chain modeling. Several comparative modeling methods, more commonly known as modeler programs, have been developed over the past several years [22,23] focusing on various parts of the problem.

As seen in the various years of CASP [24,25], the span of comparative modeling approaches [22,23] follows five basic steps: (i) selecting one or suitable templates, (ii) utilizing sensitive sequence template alignment algorithms, (iii) building a protein model using the sequence structure alignment as reference, (iv) evaluating the quality of the model, and (v) refining the model. These typical steps for the comparative modeling process are shown in Figure 1.1.

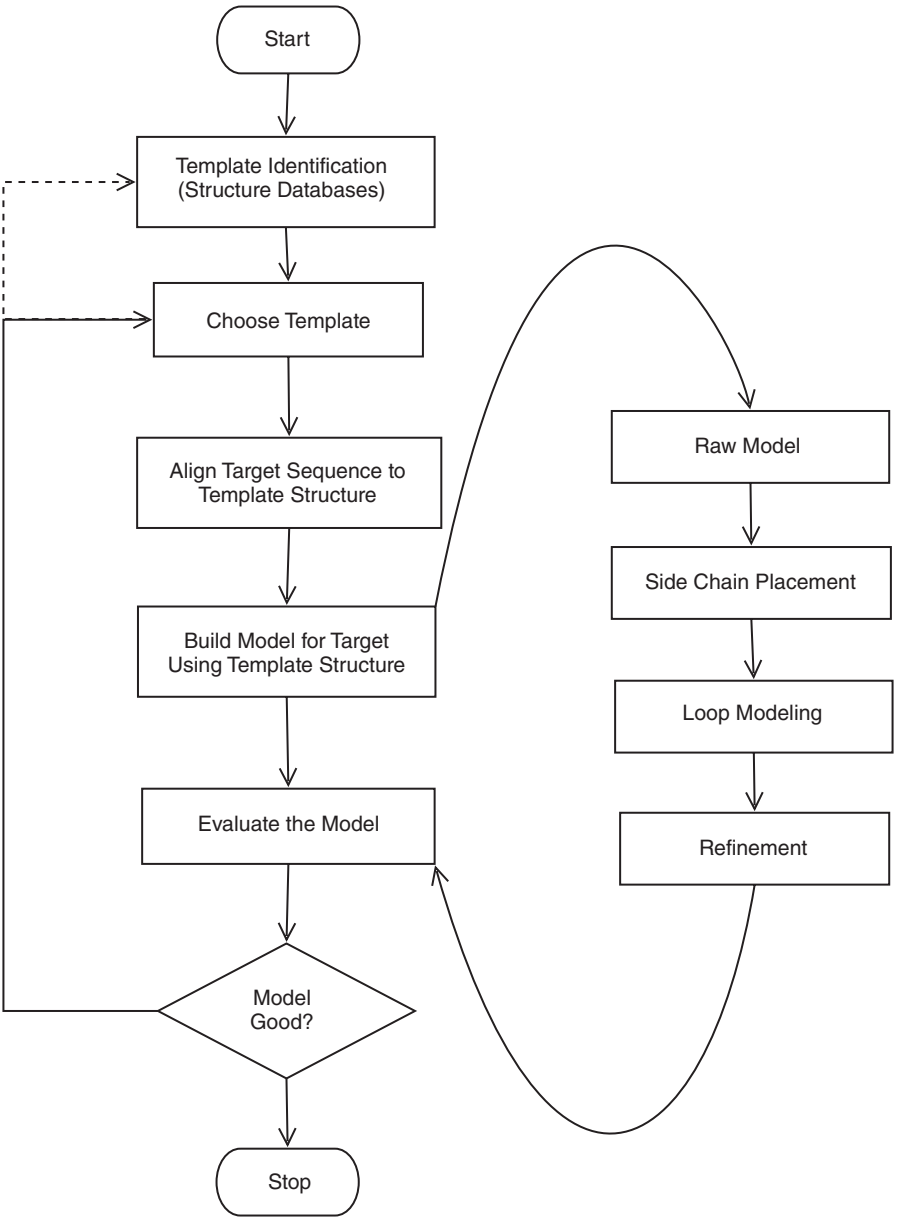

FIGURE 1.1 Flowchart for the comparative modeling process.

1.2.2. Fold Prediction (Homologous)

While satisfactory methods exist to detect homologs (proteins that share similar evolutionary ancestry) with high levels of similarity, accurately detecting homologs at low levels of sequence similarity (remote homology detection) remains a challenging problem. Some of the most popular approaches for remote homology prediction compare a protein with a collection of related proteins using methods such as Position-Specific Iterative-BLAST (PSI-BLAST) [26], protein family profiles [27], hidden Markov models (HMMs) [28,29], and Sequence Alignment and Modeling System (SAM) [30]. These schemes produce models that are generative in the sense that they build a model for a set of related proteins and then check to see how well this model explains a candidate protein.

In recent years, the performance of remote homology detection has been further improved through the use of methods that explicitly model the differences between the various protein families (classes) by building discriminative models. In particular, a number of different methods that use Support Vector Machines (SVM) [31] have been developed to produce results that are generally superior to those produced by either pairwise sequence comparisons or approaches based on generative models—provided there are sufficient training data [32–39].

1.2.3. Fold Prediction (Analogous)

Occasionally a query sequence will have a native fold similar to another known fold in a database, but the two sequences will have no detectable similarity. In many cases the two proteins will lack an evolutionary relationship as well. As the definition of this problem relies on the inability of current methods to detect sequential similarity, the set of proteins falling into this category remains in flux. As new methods continue to improve at finding sequential similarities as a result of increasing database size and better techniques, the number of proteins in question decreases. Techniques to find structures for such query sequences revolve around mounting the query sequence on a series of template structures in a process known as threading [40–42]. An objective energy function provides a score for each alignment, and the highest scoring template is chosen.

Obviously, if the correct template does not exist in the series then the method will not produce an accurate prediction. As a result of this limitation, predicting the structure of proteins in this category is as challenging as predicting protein targets that are part of the new or rare folds.

1.2.4. *Ab Initio*

Techniques to predict novel protein structure have come a long way in recent years, although a definitive solution to the problem remains elusive. Research

in this area can be roughly divided into fragment assembly [43–45] and first principle-based approaches, although occasionally the two are combined [46]. The former attempt to assign a fragment with known structure to a section of the unknown query sequence. The latter start with an unfolded conformation, usually surrounded by solvent, and allow simulated physical forces to fold the protein as would normally happen *in vivo*. Usually, algorithms from either class will use reduced representations of query proteins during initial stages to reduce the overall complexity of the problem.

Even in case of these *ab initio* prediction methods, the state-of-the-art methods [46–48] determine several template structures (using the template selection methods used in comparative modeling methods). The final protein is modeled using an assembly of fragments or substructures fitted together using a highly optimized approximate energy and statistics-based potential function.

This book presents methods developed for protein structure prediction. In particular methods and problems that are prevalent in a biennial structure prediction competition (CASP) are discussed in the first half of the book. The second half of the book discusses approaches that combine experimental and computational approaches for structure prediction and also new techniques for predicting structures of transmembrane proteins. Finally, the book discusses the applications of protein structure within the context of function prediction and drug discovery.

REFERENCES

1. G. Pandey, V. Kumar, and M. Steinbach. Computational approaches for protein function prediction: A survey. Technical Report 06-23, Department of Computer Science and Engineering, University of Minnesota, 2006.
2. D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nature Reviews. Molecular Cell Biology*, 8(12):995–1005, 2007.
3. J.C. Whisstock and A.M. Lesk. Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics*, 36(3):307–340, 2003.
4. D. Devos and A. Valencia. Practical limits of function prediction. *Proteins*, 41(1):98–107, 2000.
5. W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.
6. UniProt Consortium. The universal protein resource (uniprot). *Nucleic Acids Research*, 36(Database issue):D190–D195, 2008.
7. H.M. Berman, T.N. Bhat, P.E. Bourne, Z. Feng, G.G.H. Weissig, and J. Westbrook. The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology*, 7:957–959, 2000.
8. A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. Scop: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

9. C.A. Orengo, A.D. Mitchie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thorton. Cath- a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.
10. L. Holm and C. Sander. The fssp database: Fold classification based on structure-estructure alignment of proteins. *Nucleic Acids Research*, 24(1):206–209, 1996.
11. S. Jones, M. Stewart, A. Michie, M.B. Swindells, C. Orengo, and J.M. Thornton. Domain assignment for protein structures using a consensus approach: Characterization and analysis. *Protein Science*, 7(2):233–242, 1998.
12. W.R. Taylor and A.C. Orengo. Protein structure alignment. *Journal of Molecular Biology*, 208(1):1–22, 1989.
13. L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233(1):123–138, 1993.
14. C. Hadley and D. Jones. A systematic comparison of protein structure classifications: Scop, cath and fssp. *Structure*, 7(9):1099–1112, 1999.
15. R. Day, D.A.C. Beck, R.S. Armen, and V. Daggett. A consensus view of fold space: Combining SCOP, CATH, and the Dali Dom ain Dictionary. *Protein Science*, 12(10):2150–2160, 2003.
16. S.E. Brenner, P. Koehl, and M. Levitt. The astral compendium for sequence and structure analysis. *Nucleic Acids Research*, 28:254–256, 2000.
17. J.-M. Chandonia, N.S. Walker, L.L. Conte, P. Koehl, M. Levitt, and S.E. Brenner. ASTRAL compendium enhancements. *Nucleic Acids Research*, 30(1):260–263, 2002.
18. J.M. Chandonia, G. Hon, N.S. Walker, L.L. Conte, P. Koehl, M. Levitt, and S.E. Brenner. The astral compendium in 2004. *Nucleic Acids Research*, 32:D189–D192, 2004.
19. S.F. Altschul, W. Gish, E.W. Miller, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
20. P. Bourne and H. Weissig. *Structural Bioinformatics*. Hoboken, NJ: John Wiley & Sons, 2003.
21. T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
22. P.A. Bates and M.J.E Sternberg. Model building by comparison at casp3: Using expert knowledge and computer automation. *Proteins: Structure, Functions, and Genetics*, 3:47–54, 1999.
23. A. Fiser, R.K. Do, and A. Sali. Modeling of loops in protein structures. *Protein Science*, 9:1753–1773, 2000.
24. C. Venclovas. Comparative modeling in casp5: Progress is evident, but alignment errors remain a significant hindrance. *Proteins: Structure, Function, and Genetics*, 53:380–388, 2003.
25. C. Venclovas and M. Margelevicius. Comparative modeling in casp6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins: Structure, Function, and Bioinformatics*, 7:99–105, 2005.
26. S.F. Altschul, L.T. Madden, A.A. SchÃd'ffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

27. M. Gribskov, A.D. McLachlan, and D. Eisenberg. Profile analysis: Detection of distantly related proteins. *PNAS*, 84:4355–4358, 1987.

28. A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

29. P. Baldi, Y. Chauvin, T. Hunkapiller, and M. McClure. Hidden Markov models of biological primary sequence information. *PNAS*, 91:1053–1063, 1994.

30. K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

31. V. Vapnik. *Statistical Learning Theory*. New York: John Wiley, 1998.

32. T. Jaakkola, M. Diekhans, and D. Hassler. A dscriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1/2):95–114, 2000.

33. L. Liao and W.S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *Proceedings of the International Conference on Research in Computational Molecular Biology*, 225–232, 2002.

34. C. Leslie, E. Eskin, and W.S. Noble. The spectrum kernel: A string kernel for svm protein classification. *Proceedings of the Pacific Symposium on Biocomputing*, 564–575, 2002.

35. C. Leslie, E. Eskin, W.S. Noble, and J. Weston. Mismatch string kernels for svm protein classification. *Advances in Neural Information Processing Systems*, 20(4):467–476, 2003.

36. Y. Hou, W. Hsu, M.L. Lee, and C. Bystroff. Efficient remote homology detection using local structure. *Bioinformatics*, 19(17):2294–2301, 2003.

37. Y. Hou, W. Hsu, M.L. Lee, and C. Bystroff. Remote homology detection using local sequence-structure correlations. *Proteins: Structure, Function, and Bioinformatics*, 57:518–530, 2004.

38. H. Saigo, J.P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.

39. R. Kuang, E. Ie, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote homology detection and motif extraction. *Journal of Bioinformatics and Computational Biology*, 3:152–160, 2004.

40. D.T. Jones, W.R. Taylor, and J.M. Thorton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

41. D.T. Jones. Genthreader: An efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, 287(4):797–815, 1999.

42. J.U. Bowie, R. Luethy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:797–815, 1991.

43. K.T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268:209–225, 1997.

44. K. Karplus, R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, and R. Hughey. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins: Structure, Function, and Genetics*, 53:491–496, 2003.

45. J. Lee, S.-Y. Kim, K. Joo, I. Kim, and J. Lee. Prediction of protein tertiary structure using profesy, a novel method based on fragment assembly and conformational space annealing. *Proteins: Structure, Function, and Bioinformatics*, 56:704–714, 2004.
46. C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods in Enzymology*, 383:66–93, 2004.
47. Y. Zhang. I-tasser server for protein 3d structure prediction. *BMC Bioinformatics*, 9:40, 2008.
48. Y. Zhang, A.J. Arakaki, and J. Skolnick. Tasser: An automated method for the prediction of protein tertiary structures in casp6. *Proteins: Structure, Function, and Bioinformatics*, 7:91–98, 2005.