
PART IV

PATTERN RECOGNITION IN QUATERNARY STRUCTURES

PREDICTION OF PROTEIN QUATERNARY STRUCTURES

Akbar Vaseghi¹, Maryam Faridounnia², Soheila Shokrollahzade³,
Samad Jahandideh⁴, and Kuo-Chen Chou^{5,6}

¹*Department of Genetics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran*

²*Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands*

³*Department of Medicinal Biotechnology, Iran University of Medical Sciences, Tehran, Iran*

⁴*Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research Institute, La Jolla, CA, USA*

⁵*Department of Computational Biology, Gordon Life Science Institute, Belmont, MA, USA*

⁶*Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia*

14.1 INTRODUCTION

In all living organisms, a protein sequence built up by 20 main amino acid residues and linked through peptide bonds, is encoded by genes and is called the *primary structure*. Depending on the sequence of amino acids, environmental conditions, and a number of other factors, the polypeptide chain folds into very well-defined conformational segments called the *secondary structure* (β -strands, α -helices, etc.). These segments fold into larger conformations with different arrangements creating a *tertiary structure*. In order to be functional and stable, many of these polypeptide chains

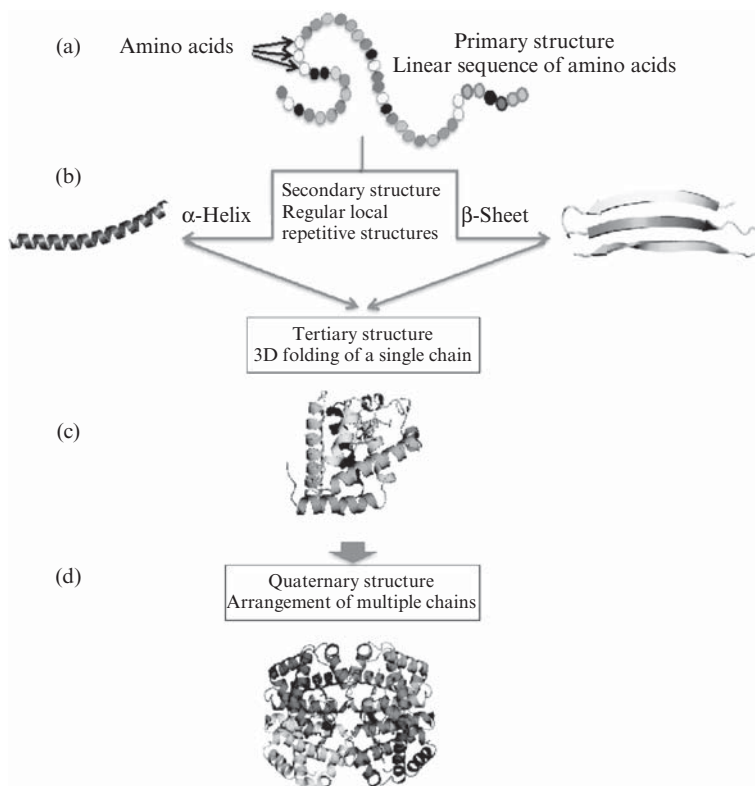


Figure 14.1 The different protein structures: (a) primary structure (polypeptide chain), (b) secondary structure (α -helix), (c) tertiary structure (example: myoglobin), and (d) quaternary structure (example: hemoglobin).

with tertiary structures assemble together as *quaternary structures* (Figure 14.1). These protein assemblies are involved in a variety of biological processes such as cellular metabolism, signal transduction, chromosome transcription, replication, and DNA damage repair.

Quaternary structures are the spatial arrangements including multiple copies of one or multiple types of polypeptide chains, assembling through noncovalent interactions to enable a biological function, which depending on the conditions can be very specific or multipurpose. This creates the protein universe with various classes of subunit constructions, such as monomer, dimer, trimer, tetramer, pentamer, and hexamer [22] (Figure 14.2).

Knowledge of the protein quaternary structure is important because it enables to discover the biological function of the protein and, hence, it enables to target this function during drug development [18]. For instance, the hemoglobin [54], potassium channel [19, 30], and the M2 proton channel [69] are tetramers; while the phospholamban [53], Gamma-AminoButyric Acid type A (GABAA) receptor [20, 73]), and

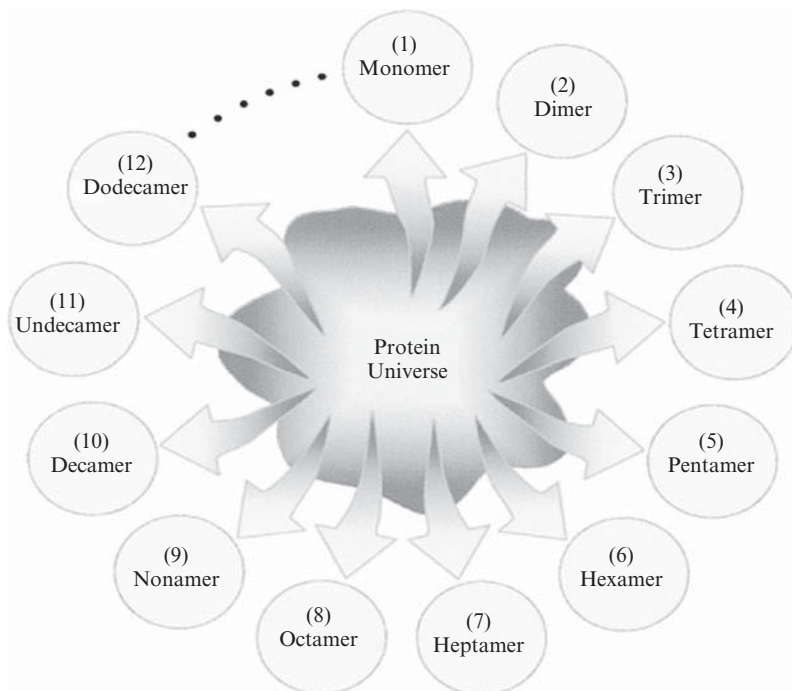


Figure 14.2 The schematic representation of different polypeptide chains that form various oligomers. *Source:* Chou and Cai (2003) [22]. Reproduced with permission of John Wiley and Sons, Inc.

alpha7 nicotinic acetylcholine receptor [20] are pentamers. The sodium channel is a monomer [11], whereas the p7 hepatitis C virus is a hexamer [52].

Among the aforementioned proteins, hemoglobin is a classical example that is made up of two α -chains and two β -chains, and these four chains aggregate into one structure to perform the protein's cooperative function during the oxygen-transportation process, as elucidated from the viewpoint of low-frequency collective motion [13–15, 54]. In addition, recent findings have revealed that the novel molecular wedge allosteric drug-inhibited mechanism [56] for the M2 proton channel can be understood [39] through a unique left-handed twisting packing [23] arrangement of four transmembrane helices from four identical protein chains [69].

There are three categories of rotational symmetry for oligomeric proteins, including (i) cyclic symmetry C_n , which has an n -fold symmetry by $360^\circ/n$ rotations ($n = 2, 3, 4, \dots$); for example, Figure 14.3a shows an object which has C_5 symmetry with a fivefold rotational axis. (ii) Dihedral symmetry, D_n , is generated when an n -fold symmetry intersects with a twofold axis of symmetry. For example, Figure 14.3b shows D_4 symmetry. (iii) Polyhedral symmetries such as cubic, tetrahedrons, and icosahedrons have 12, 24, or 60 identical subunits (Figure 14.3c, e, and d). The most common symmetry for soluble proteins is C_2 in homodimers. Among other common symmetries, D_2 tetramers are more common than C_4 (potassium

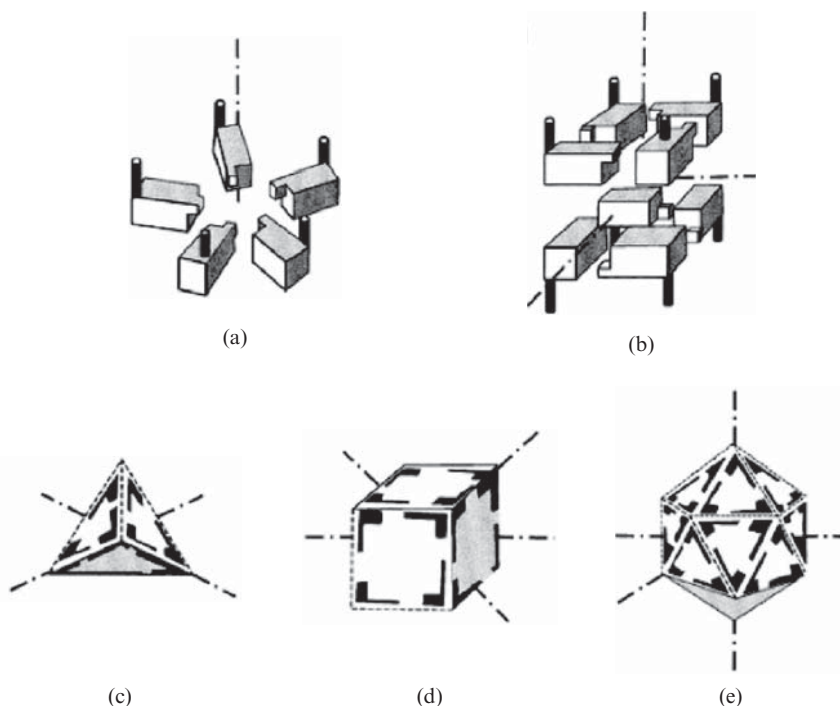


Figure 14.3 A schematic drawing to illustrate protein oligomers with: (a) C_5 symmetry, (b) D_4 symmetry, (c) tetrahedron symmetry, (d) cubic symmetry, and (e) icosahedrons symmetry. Source: Chou and Cai (2003) [22]. Reproduced with permission of John Wiley and Sons, Inc.

channel) and C_5 (pentameric acetylcholine receptor), and D_3 (of bacteriorhodopsin) hexamers are more common than C_6 ([40]).

Determination of the protein structure and understanding its function is essential for any relevant medical, engineering, or pharmaceutical applications. Therefore, the study of quaternary structure of proteins, despite all the obstacles in acquiring data from large macromolecular assemblies, is one of the major goals in biomolecular sciences. Although the structure of a large number of proteins is experimentally solved, there exist many proteins with unknown fold and without any obvious homology. The protein structure can be investigated by experimental [4, 43, 45, 71, 75, 79] and computational methods, namely, Comparative Modeling (CM) [1, 2, 57, 60, 67, 83]. Considering the fact that determination of protein structure using the experimental methods is expensive, labor intensive, and time consuming, the Structural Genomics (SG) project has been introduced with the goal of developing and integrating new technologies that cover the gene-to-structure process. This SG approach facilitates high-throughput structural biology. However, regarding the considerably higher speed and lower cost of computing, the development of computational methods for protein structure prediction is very promising for prediction of unsolved protein structures [65].

The rest of this chapter is organized as follows. In Sections 14.2–14.5, we discuss respectively protein structure prediction, template-based predictions, critical assessment of protein structure prediction, and quaternary structure prediction. Finally, in Section 14.6, we present the conclusion of this chapter.

14.2 PROTEIN STRUCTURE PREDICTION

Development of large-scale sequencing techniques speeded up sequencing and resulted in a huge gap between the number of known protein sequences and the number of solved structures (Figure 14.4). Accurate prediction of protein structure is a difficult task due to (i) limited knowledge about stability of protein structure, (ii) the role of chaperons in the folding process of proteins from different families, (iii) multiple folding pathways to attain the native state being evidenced for a single protein, and (iv) an enormous quantity of possible conformations that can be suggested for each protein. Nevertheless, the goal of protein structure prediction projects is to reduce this gap. During its early stage, the problem was simplified by mimicking one possible mechanism of protein folding, which includes the two folding steps: (i) formation of local secondary structures and (ii) arrangement of secondary structures to achieve the folded conformation. Later, before trying to tackle tertiary structure prediction, for decades, many secondary structure prediction methods were developed. Here, we briefly introduce the three generations of secondary structure prediction methods and continue with popular methods for prediction of tertiary structure.

14.2.1 Secondary Structure Prediction

Up to now, different methods have been developed for prediction of secondary structure. Here, we introduce three different generations of these methods in separate sections.

14.2.1.1 First Generation. In the first generation, a few popular methods were developed by simple analysis of amino acids distribution in α -helices and β -strands. The first method that was widely used for prediction of secondary structures was the Chou–Fasman method [24]. This method was developed in 1974 on the basis of statistical analysis of preference and avoidance of all 20 amino acids in α -helices and β -strands. They simply calculated the propensity of all 20 amino acids for α -helices and β -strands in a small database of solved protein structures. The Chou–Fasman method consists of a number of simple rules. For example, if the property in a window of six amino acids (for a helix) or five amino acids (for a strand) is above a predefined threshold value, this segment is the nucleation point of a possible secondary structure.

The next standard method was the Garnier, Osuguthorpe, and Robson (GOR) method [33]. The GOR method, developed four years after the Chou–Fasman method, was slightly more sophisticated and more accurate. Secondary structure prediction was primarily based on a sliding window of 17 residues and assigning a value to each residue that expresses the likelihood of it being in a particular secondary structure.

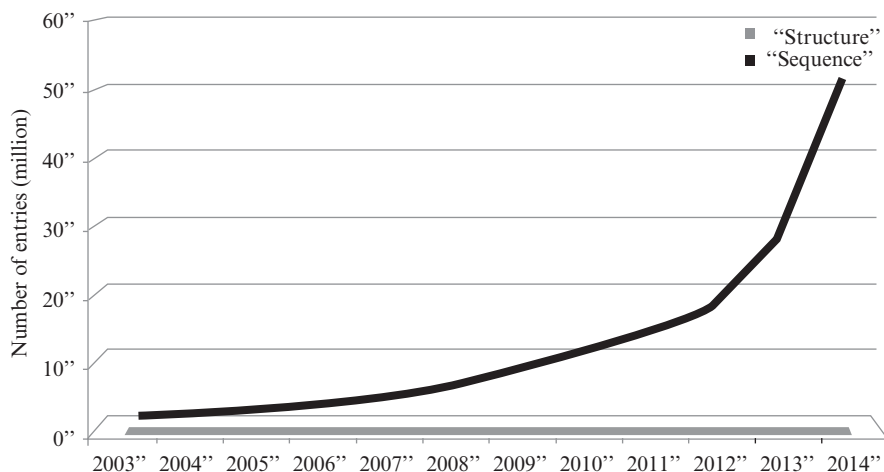


Figure 14.4 Comparison of determined number of protein sequences and protein structures based on statistical data from PDB and UniProtKB.

14.2.1.2 Second Generation. The second generation of secondary structure prediction methods covers a combination of methodology development and feature generation. Different predictor methods including statistical learning and machine learning are applied for prediction of secondary structures [62, 63, 72]. Furthermore, a class of features including physicochemical properties and sequence patterns was defined to fit into the predictor models [34]. In this generation, accuracy of secondary structure prediction had 10% improvement compared to the methods in the first generation with an accuracy of 50–60%.

14.2.1.3 Third Generation. These methods outperformed developed methods in the earlier generations in terms of accuracy [10, 64]. Their accuracy reached close to 80% [55]. The main cause for improvement in accuracy was adding evolutionary information to the list of features generated from multiple sequence alignments. For example, Rost and Sander in 1993 developed the PHD method that is composed of several cascading neural networks [61]. In the neural network, homologous sequences that are determined by BLAST [55] are aligned using the MaxHom alignment [68] of known structures; then, generated conservation scores are used to train the network, which then can be used to predict the secondary structure of the aligned sequences of the unknown protein.

In addition to prediction of secondary structure, many other methods are developed for prediction of local attributes of proteins, such as disordered regions, membrane-spanning beta-barrels, and transmembrane helices.

14.2.2 Modeling of Tertiary Structure

Despite the advances in protein structure determination in recent decades, some classes of proteins such as membrane proteins, that is, proteins integrated into a

complex lipid bilayer environment—that are of eminent importance in biological sciences and medicinal biology—and other proteins with large quaternary structure arrangements, are not easy to crystallize or are too large for determination with Nuclear Magnetic Resonance (NMR) [12, 44]. Therefore, developing accurate protein structure prediction tools is very important, as many computational methods that have been developed until now comprise machine-learning-based methods [2], SCWRL and MolIDE [74], Rosetta [1, 60], MODELLER [67], homology modeling, fold recognition, threading, and *ab initio* prediction. In this section, we briefly introduce more popular methodologies for tertiary structure prediction.

14.3 TEMPLATE-BASED PREDICTIONS

Homology modeling and threading methods are two types of template-based approaches. The homology modeling method needs to have the homologous protein structure as template and threading methods are a new approach in fold recognition, in which the tool attempts to fit the sequence in the known structures.

14.3.1 Homology Modeling

Considering that high sequence similarity results in similar three-dimensional (3D) structures, a large number of methods (homology modeling, comparative modeling, or template-based modeling) use a properly known structure as a seeding point. This type of method has five stages, including (i) finding a proper template, (ii) multiple sequence alignment of the query protein to the templates, (iii) modeling the target structure including main chain modeling, loop modeling, and side chain modeling, (iv) energy refinement, and (v) model evaluation. *Swiss-model* [36] and *3D Jigsaw* [3] are some of the most commonly used servers for homology modeling.

14.3.2 Threading Methods

The threading method is a protein structure prediction technique applied when there is not enough sequence similarity between the target and the template. In this method, the target sequence is pasted onto a known fold and the compatibility of the sequence and the fold is evaluated by calculating a score. The local secondary structure, the environment, and the pairwise interaction of side chains of close amino acids are the structural properties that are used to evaluate the fit. FFAS03 [41], 3DPSSM [46], and I-TASSER [66] are the most popular threading programs.

14.3.3 *Ab initio* Modeling

Anfinsen's hypothesis implies that the native state of the protein represents the global free energy minimum. Considering this hypothesis, *ab initio* methods try to find these global minima of the protein [29] and predict tertiary structure without reference to a specific template with homologous structure [51]. Finding the correct conformation

using *ab initio* modeling requires an efficient search method for exploring the conformational space to find the energy minima and an accurate potential function that calculates the free energy of a given structure. *Rosetta* is the most popular program in this category [42].

14.4 CRITICAL ASSESSMENT OF PROTEIN STRUCTURE PREDICTION

The best method for evaluation of the performance of structure prediction methods is *in silico* test on unknown structures, which is the focus of the Critical Assessment of protein Structure Prediction (CASP) meetings [50]. CASP invites the scientists that developed methods for protein structure prediction to model structure of proteins with solved structure but not released yet. The models are submitted and compared to the experimentally determined structures. The most accurate methods are introduced and the progression of the field is discussed during these biannual meetings. Evaluation of the results is carried out in different categories including tertiary structure prediction, residue-residue contact prediction, disordered regions prediction, function prediction, model quality assessment, model refinement, and high-accuracy template-based prediction.

14.5 QUATERNARY STRUCTURE PREDICTION

The Protein Data Bank (PDB) contains more than 100,000 protein structures mostly determined by X-ray crystallography and NMR [5]. By convention, a crystallographic PDB entry reports atomic coordinates for the crystal ASymmetric Unit (ASU). ASU includes the subunit contacts and intermolecular contacts that define the quaternary structure. In addition to PDB, many databases exist such as Protein Biological unit Database (ProtBuD) [78], Protein InTerfaces and Assemblies (PITA) [58], Probable Quaternary Structure (PQS) [37], Protein Quaternary Structure investigation) PiQSi [48], and Protein Interfaces, Surfaces, and Assemblies (PISA) [47] designed for quaternary structure studies. PQS, PISA, and PITA are developed at the European Bioinformatics Institute (EBI-EMBL, Hinxton, UK). These databases also include the ones that are not limited to biological units. PITA and PQS apply crystal symmetries to the molecules in the ASU, select neighbors, and score each pairwise interface on the basis of the buried area, plus a statistical potential in PITA or a solvation energy term in PQS [6]. Nevertheless, PiQSi facilitated the visual inspection of the quaternary structure of protein complexes in PDB [49]. PiQSi characterizes a query protein as belonging to monomer, homo-oligomer, or hetero-oligomer according to its sequence information alone. The first version of PiQSi annotated over 10,000 structures from the PDB biological unit and corrected the quaternary state of approximately 15% of them. All data are also available at <http://www.PiQSi.org> [49].

Using the above-mentioned databases, a few sequence-based computational methods have been developed for the prediction of protein quaternary structure

using statistical models or machine learning methods. The earliest method for prediction of quaternary structure was a decision tree model that uses simple sequence-based features to discriminate between the primary sequences of homodimers and non-homodimers [31]. Subsequently, new algorithms were developed for the prediction of different types of homomeric structures ([16], Chou, 2004, [21]) and heteromeric structures [9].

Features, databases, and prediction models are three important elements in developing a reliable and accurate prediction methodology. Different combinations of predictor models and features have been examined for quaternary structure prediction. For example, pseudo amino acid composition, originally introduced by Kou-Chen Chou at The Gordon Life Science Institute, was a feature for improving the prediction of protein subcellular location [17] (Figure 14.5). Pseudo amino acid composition was introduced to get the desired results when trying sequence-similarity-search-based prediction. This kind of approach fails when a query protein does not have significant homology to the known protein(s). Thus, various discrete models were proposed that do not rely on sequence-order (see Reference [17] for more details). Later, pseudo amino acid composition was used for the prediction of quaternary structure [17, 22]. Moreover, the functional domain composition information, which has been very useful in identifying enzyme functional classes [25] and protease types [26], was fed into the Nearest Neighbor (NN) algorithm [28] that is proved to be very successful in prediction of protein subcellular localization [26] as prediction engine for prediction of quaternary structure [76].

Training of models using small database selected from the population increases the error rate in testing. In such circumstances, predictor performance can be improved by

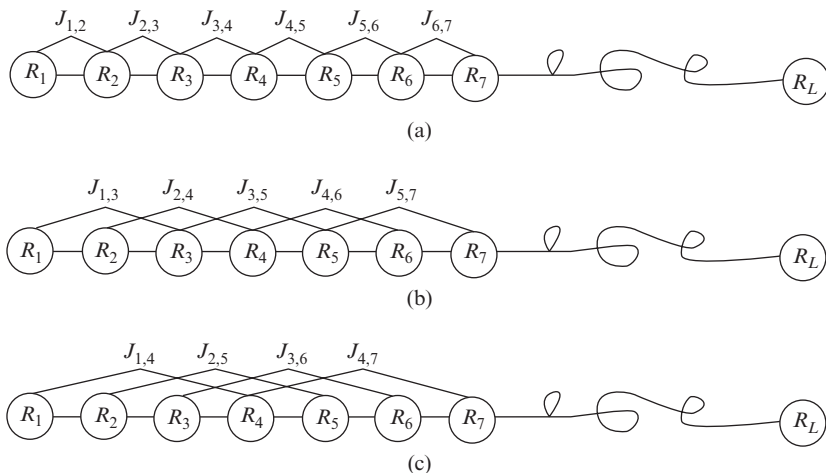


Figure 14.5 Definition procedure for the huge number of possible sequence order patterns as a set of discrete numbers. Panels (a–c) represent the correlation mode between all the most contiguous residues, all the second-most contiguous residues, and all the third-most contiguous residues, respectively. Source: Chou and Cai (2003) [22]. Reproduced with permission of John Wiley and Sons, Inc.

using larger a data set in the training step. Moreover, redundancy of protein sequences can result in overfitting. In this case, overfitting refers to receiving higher weights by nonrelevant features that are redundantly repeated in the data set. However, there is overfitting in the use of features, models, or evaluation procedures that violate parsimony, that is, they use more complicated approaches than are necessary. A general rule is that the best predictor model is the simplest. For example, if a simple model, for example, the linear regression model, perfectly fits to a database, then applying an overfitting-prone complicated model like an artificial neural network results in adding a level of complexity without any corresponding benefit in performance or, even worse, with poorer performance than the simpler model.

After the first publication on classification of protein quaternary structure, the Support Vector Machines (SVMs) were introduced to classify quaternary structure properties from the protein primary sequences [81]. SVM is a new type of data mining method applied with success to different problems in bioinformatics and computational biology, such as translation initiation sites [84], membrane protein types [8], protein–protein interactions [8], and protein subcellular localization [38]. In this study, a binary SVM was applied to discriminate between the primary sequences of homodimers and non-homodimers, and the obtained results were similar to the previous investigation by Garian in terms of performance [32]. Subsequently, different methodologies were applied for classification and prediction of protein quaternary structure—for example, the NN algorithm [80], threading [35], and the Function for the Degree Of Disagreement (FDOD) are applied to discriminate between homodimers and other homo-oligomeric proteins from the primary structure [70]. By extending the problem to different types of homo-oligomer and hetero-oligomer quaternary structures, recently Quat-2L was developed for predicting protein quaternary structure and identifying the query protein as monomer, homo-oligomer, or hetero-oligomer for the first step (with a success rate of 71.14%), and for different types of homo-oligomers and hetero-oligomers for the second step (with success rates of 76.91% and 82.52%, respectively). Quat-2L is available as a web-server at <http://icpr.jci.jx.cn/bio-info/Quat-2L> [77]. Hybridization of pseudo amino acids composition and Quat-2L mode can be used to predict protein quaternary structural attribute of a protein chain according to its sequence information alone with higher success rate [77]. The sequence-segmented pseudo amino acids composition approach can capture essential information about the compositions and hydrophobicity of residues in the surface patches buried in the interfaces of associated subunits [82].

Three conventional procedures for performance evaluation of predictor methods are defined as self-consistency, cross-validation, and jackknife [30]. The self-consistency procedure is as simple as training and testing of the predictor model on the same data set that only evaluates correlation between input features and outputs. On the other hand, during cross-validation and jackknife tests, none of the test samples is presented in training procedures. The jackknife test is a type of cross-validation in which only one sample is fed into the model as testing sample after each training procedure; therefore, the number of iterations is equal to the total number of samples in data set. However, in the cross-validation procedure, the data

set is divided into some subsets and for each iteration one subset is eliminated during training and the total number of iteration is equal to the number of subsets [76].

14.6 CONCLUSION

The function of a protein is related to its quaternary structure and thus prediction of quaternary structure from the protein sequence is very useful and can be obtained by studying the relation of the quaternary structural attribute of a protein chain and its sequence feature. Predicting quaternary structure of a protein is a difficult task. However, the development and growth of novel statistical learning and machine learning methods, the increase in the number of solved structures, and the improvement of computational resources, have together increased the chances of development of more accurate prediction methods in the near future. Among protein structure prediction problems, prediction of quaternary structure is fairly new and the methods that have been developed are still immature. More efforts are needed to increase the accuracy of these methods to a higher level of reliability to become applicable for biologists. We are very optimistic that accuracy of quaternary structure prediction will be improved by using recently developed state-of-the-art machine learning methods, such as the random forest method [7] and multi-class SVM [59].

ACKNOWLEDGMENTS

We express our thanks to all members of “The First Protein Structural Bioinformatics Online Workshop.” This chapter is extracted from our literature review and discussion on protein structure prediction in this workshop.

REFERENCES

1. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
2. Baldi P. *Bioinformatics: The Machine Learning Approach*. The MIT Press; 2001.
3. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* 2001;5:39–46.
4. Benesch JL, Robinson CV. Mass spectrometry of macromolecular assemblies: preservation and dissociation. *Current Opin Struct Biol* 2006;16:245–251.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000;28:235–242.
6. Berman HM, Battistuz T, Bhat T, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S. The protein data bank. *Acta Crystallogr Sect D: Biol Crystallogr* 2002;58:899–907.
7. Breiman L. Random forests. *Mach Learn* 2001;45:5–32.

8. Cai YD, Liu XJ, Xu XB, Chou KC. Support vector machines for predicting membrane protein types by incorporating quasi-sequence-order effect. *Internet Electron J Mol Des* 2002;1:219–226.
9. Carugo O. A structural proteomics filter: prediction of the quaternary structural type of hetero-oligomeric proteins on the basis of their sequences. *J Appl Crystallogr* 2007;40:986–989.
10. Chandonia JM, Karplus M. New methods for accurate prediction of protein secondary structure. *Proteins* 1999;35:293–306.
11. Chen Z, Alcayaga C, Suarez-Isla BA, O'Rourke B, Tomaselli G, Marban E. A “minimal” sodium channel construct 2002.
12. Cheng J, Tegge AN, Baldi P. Machine learning methods for protein structure prediction. *Biomed Eng IEEE Rev* 2008;1:41–49.
13. Chou K-C. The biological functions of low-frequency phonons. 4. Resonance effects and allosteric transition. *Biophys Chem* 1984;20:61–71.
14. Chou K-C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys Chem* 1988;30:3–48.
15. Chou K-C. Low-frequency resonance and cooperativity of hemoglobin. *Trends Biochem Sci* 1989;14:212–213.
16. Chou K-C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophys Res Commun* 2000;278:477–483.
17. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 2001;43:246–255.
18. Chou K-C. Structural bioinformatics and its impact to biomedical science. *Curr Med Chem* 2004;11:2105–2134.
19. Chou K-C. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. *Biochem Biophys Res Commun* 2004;316:636–642.
20. Chou K-C. Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. *Biochem Biophys Res Commun* 2004;319:433–438.
21. Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 2005;21:10–19.
22. Chou KC, Cai YD. Predicting protein quaternary structure by pseudo amino acid composition. *Proteins* 2003;53:282–289.
23. Chou KC, Maggiora GM, Nemethy G, Scheraga HA. Energetics of the structure of the four-alpha-helix bundle in proteins. *Proceedings of the National Academy of Sciences of the United States of America (PNAS USA)* 1988;85:4295–4299.
24. Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry* 1974;13:222–245.
25. Chou K-C, Shen H-B. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J Proteome Res* 2007;6:1728–1734.
26. Chou K-C, Shen H-B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem Biophys Res Commun* 2008;376:321–325.
27. Chou KC, Zhang CT. Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 1995;30:275–349.

28. Cover T, Hart P. The nearest neighbor decision rule. *IEEE Trans Inform Theory* 1967;13:21–27.
29. Defay T, Cohen FE. Evaluation of current techniques for ab initio protein structure prediction. *Proteins* 1995;23:431–445.
30. Doyle DA, Morais CJ, Pfuertner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R. The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* 1998;280:69–77.
31. Garian R. Prediction of quaternary structure from primary structure. *Bioinformatics* 2001;17:551–556.
32. Garian R. Prediction of quaternary structure from primary structure. *Bioinformatics* 2001;17:551–556.
33. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* 1996;266:540–553.
34. Geourjon C, Deleage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci* 1995;11:681–684.
35. Grimm V, Zhang Y, Skolnick J. Benchmarking of dimeric threading and structure refinement. *Proteins* 2006;63:457–465.
36. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb viewer: an environment for comparative protein modeling. *Electrophoresis* 1997;18(15):2714–2723.
37. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. *Trends Biochem Sci* 1998;23:358–361.
38. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 2001;17:721–728.
39. Huang RB, Du DS, Wang CH, Chou KC. An in-depth analysis of the biological functional studies based on the NMR M2 channel structure of influenza A virus. *Biochem Biophys Res Commun* 2008;377:1243–1247.
40. Janin J, Bahadur RP, Chakrabarti P. Protein–protein interaction and quaternary structure. *Q Rev Biophys* 2008;41:133–180.
41. Jaroszewski L, Rychlewski L, Li Z, Li WZ, Godzik A. FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res* 2005;33:W284–W288.
42. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. *Proteins* 2007;69:57–67.
43. Blundell TL, Johnson LN. *Protein Crystallography*. New York: Academic; 1976.
44. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 1999;287:797–815.
45. Karlsson R. SPR for molecular interaction analysis: a review of emerging application areas. *J Mol Recogn* 2004;17:151–161.
46. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM1. *J Mol Biol* 2000;299(2):501–522.
47. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 2007;372:774–797.
48. Levy ED. PiQSi: protein quaternary structure investigation. *Structure* 2007;15:1364–1367.
49. Levy ED. PiQSi: protein quaternary structure investigation. *Structure* 2007b;15:1364–1367.

50. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) – round X. *Proteins* 2014;82 (Suppl 2):1–6.
51. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins* 1999;37:177–185.
52. OuYang B, Xie S, Berardi MJ, Zhao XM, Dev J, Yu W, Sun B, Chou JJ. Unusual architecture of the p7 channel from hepatitis C virus. *Nature* 2013;498:521–525.
53. Oxenoid K, Chou JJ. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc Natl Acad Sci USA* 2005;102:10870–10875.
54. Perutz MF. The hemoglobin molecule. *Sci Am* 1964;211:65–76.
55. Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert GP, Lund O. Prediction of protein secondary structure at 80% accuracy. *Proteins* 2000;41:17–20.
56. Pielak RM, Jason R, Schnell JR, Chou JJ. Mechanism of drug inhibition and drug resistance of influenza A M2 channel. *Proc Natl Acad Sci USA* 2009;106:7379–7384.
57. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 2006;34:D291–D295.
58. Ponstingl H, Kabir T, Thornton JM. Automatic inference of protein quaternary structure from crystals. *J Appl Cryst* 2003;36:1116–1122.
59. Rashid M, Saha S, Raghava GP. Support vector machine-based method for predicting sub-cellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinf* 2007;8:337.
60. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* 2004;55:656–677.
61. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 1993a;90:7558–7562.
62. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993b;232:584–599.
63. Rost B, Sander C. Secondary structure prediction of all-helical proteins in two states. *Protein Eng* 1993c;6:831–836.
64. Rost B, Sander C. Third generation prediction of secondary structures. *Methods Mol Biol* 2000;143:71–95.
65. Rost B, Liu J, Przybylski D, Nair R, Wrzeszczynski KO, Bigelow H, Ofra Y. Prediction of protein structure through evolution. In: *Handbook of Chemoinformatics: From Data to Knowledge* 4 vols. Wiley; 2003. p 1789–1811.
66. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 2010;5:725–738.
67. Sali A, Blundell T. Comparative protein modelling by satisfaction of spatial restraints. *Protein Struct Distance Anal* 1994;64:C86.
68. Sander C, Schneider R. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 1991;9:56–68.
69. Schnell JR, Chou JJ. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* 2008;451:591–595.
70. Song J, Tang H. Accurate classification of homodimeric vs other homooligomeric proteins using a new measure of information discrepancy. *J Chem Inform Comput Sci* 2004;44:1324–1327.

71. Sund H, Weber K. The quaternary structure of proteins. *Angew Chem Int Edit Engl* 1966;5:231–245.
72. Taylor WR, Thornton JM. Prediction of super-secondary structure in proteins. *Nature* 1983;301:540–542.
73. Tretter V, Ehya N, Fuchs K, Sieghart W. Stoichiometry and assembly of a recombinant GABAA receptor subtype. *J Neurosci* 1997;17:2728–2737.
74. Wang Q, Canutescu AA, Dunbrack RL. SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat Protoc* 2008;3:1832–1847.
75. Wuthrich K. *NMR of Proteins and Nucleic Acids*. New York: Wiley; 1986.
76. Xiao X, Wang P, Chou K-C. Predicting the quaternary structure attribute of a protein by hybridizing functional domain composition and pseudo amino acid composition. *J Appl Crystallogr* 2009;42:169–173.
77. Xiao X, Wang P, Chou K-C. Quat-2L: a web-server for predicting protein quaternary structural attributes. *Mol Diversity* 2011;15:149–155.
78. Xu Q, Canutescu A, Obradovic Z, Dunbrack RL Jr. ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics* 2006;22:2876–2882.
79. Yan Y, Marriott G. Analysis of protein interactions using fluorescence technologies. *Curr Opin Chem Biol* 2003;7:635–640.
80. Yu X, Wang C, Li Y. Classification of protein quaternary structure by functional domain composition. *BMC Bioinform* 2006;7:187.
81. Zhang S-W, Pan Q, Zhang H-C, Zhang Y-L, Wang H-Y. Classification of protein quaternary structure with support vector machine. *Bioinformatics* 2003;19:2390–2396.
82. Zhang S-W, Chen W, Yang F, Pan Q. Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. *Amino Acids* 2008;35:591–598.
83. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* 2001;44:336–343.
84. Zien A, Rätsch G, Mika S, Schölkopf B, Lengauer T, Müller K-R. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* 2000;16:799–807.