Machine Learning Methods for Protein Structure Prediction

Jianlin Cheng, Allison N. Tegge, *Member, IEEE*, and Pierre Baldi, *Senior Member, IEEE*

Methodological Review

***Abstract*—Machine learning methods are widely used in bioinformatics and computational and systems biology. Here, we review the development of machine learning methods for protein structure prediction, one of the most fundamental problems in structural biology and bioinformatics. Protein structure prediction is such a complex problem that it is often decomposed and attacked at four different levels: 1-D prediction of structural features along the primary sequence of amino acids; 2-D prediction of spatial relationships between amino acids; 3-D prediction of the tertiary structure of a protein; and 4-D prediction of the quaternary structure of a multiprotein complex. A diverse set of both supervised and unsupervised machine learning methods has been applied over the years to tackle these problems and has significantly contributed to advancing the state-of-the-art of protein structure prediction. In this paper, we review the development and application of hidden Markov models, neural networks, support vector machines, Bayesian methods, and clustering methods in 1-D, 2-D, 3-D, and 4-D protein structure predictions.**

***Index Terms*—Bioinformatics, machine learning, protein folding, protein structure prediction.**

I. Introduction

A protein is a polymeric macromolecule made of amino acid building blocks arranged in a linear chain and joined together by peptide bonds. The linear polypeptide chain is called the primary structure of the protein. The primary structure is typically represented by a sequence of letters over a 20-letter alphabet associated with the 20 naturally occurring amino acids.

In its native environment, the chain of amino acids (or residues) of a protein folds into local secondary structures including alpha helices, beta strands, and nonregular coils [3], [4]. The secondary structure is specified by a sequence classifying each amino acid into the corresponding secondary structure element (e.g., alpha, beta, or gamma). The secondary structure elements are further packed to form a tertiary structure depending on hydrophobic forces and side chain interactions, such as hydrogen bonding, between amino acids [5]–[7]. The tertiary structure is described by the x, y, and z coordinates of all the atoms of a protein or, in a more coarse description, by the coordinates of the backbone atoms. Finally, several related protein chains can interact or assemble together to form protein complexes. These protein complexes correspond to the protein quaternary structure. The quaternary structure is described by the coordinates of all the atoms, or all the backbone atoms in a coarse version, associated with all the chains participating in the quaternary organization, given in the same frame of reference.

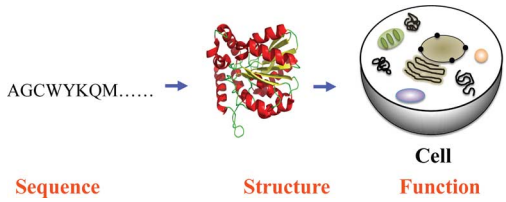

Fig. 1. Protein sequence-structure-function relationship. A protein is a linear polypeptide chain composed of 20 different kinds of amino acids represented by a sequence of letters (left). It folds into a tertiary (3-D) structure (middle) composed of three kinds of local secondary structure elements (helix – red; beta-strand – yellow; loop – green). The protein with its native 3-D structure can carry out several biological functions in the cell (right).

In a cell, proteins and protein complexes interact with each other and with other molecules (e.g., DNA, RNA, metabolites) to carry out various types of biological functions ranging from enzymatic catalysis, to gene regulation and control of growth and differentiation, to transmission of nerve impulses [8]. Extensive biochemical experiments [5], [6], [9], [10] have shown that a protein's function is determined by its structure. Thus, elucidating a protein's structure is key to understanding its function, which in turn is essential for any related biological, biotechnological, medical, or pharmaceutical applications.

Experimental approaches such as X-ray crystallography [11], [12] and nuclear magnetic resonance (NMR) spectroscopy [13], [14] are the main techniques for determining protein structures. Since the determination of the first two protein structures (myoglobin and haemoglobin) using X-ray crystallography [5], [6], the number of proteins with solved structures has increased rapidly. Currently, there are about 40 000 proteins with empirically known structures deposited in the Protein Data Bank (PDB) [15]. This growing set of solved structures

Manuscript received August 27, 2008; revised October 03, 2008. First published November 05, 2008; current version published December 12, 2008. The work of J. Cheng was supported by an MU research board grant. The work of A. N. Tegge was supported by an NLM fellowship. The work of P. Baldi was supported in part by the MU bioinformatics Consortium, in part by NIH Biomedical Informatics Training under Grant (LM-07443-01), and in part by the National Science Foundation under MRI Grant (EIA-0321390) and Grant (0513376).

J. Cheng is with the Computer Science Department, University of Missouri, Columbia, MO 65211 USA (e-mail: chengji@missouri.edu).

A. N. Tegge is with the Informatics Institute, University of Missouri, Columbia, MO 65211 USA (e-mail: ategge@mizzou.edu).

P. Baldi is with the Department of Computer Science and the Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697 USA(e-mail: pfbaldi@uci.edu).

Digital Object Identifier 10.1109/RBME.2008.2008239

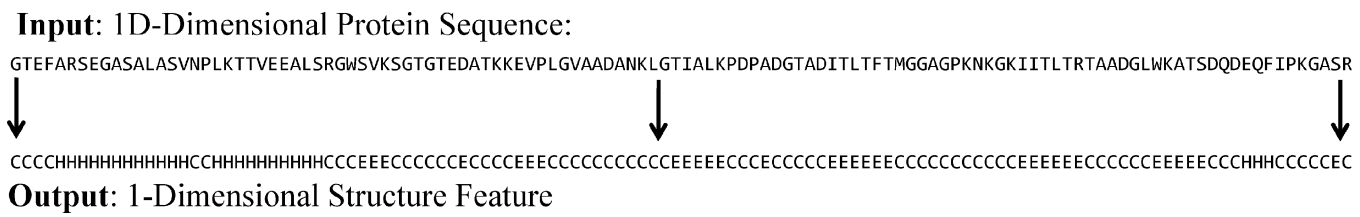

Fig. 2. One-dimensional protein structure prediction. Three-dimensional example of 1-D protein structure prediction where the input primary sequence of amino acid is "translated" into an output sequence of secondary structure assignments for each amino acid (C = coil; H = helix; E = beta-strand [extended sheet)].

provides invaluable information to help further understand how a protein chain folds into its unique 3-D structure, how chains interact in quaternary complexes, and how to predict structures from primary sequences [16].

Since the pioneering experiments [1], [2], [5], [6], [17] showing that a protein's structure is dictated by its sequence, predicting protein structure from its sequence has become one of the most fundamental problems in structural biology (Fig. 1). This is not only a fundamental theoretical challenge but also a practical one due to the discrepancy between the number of protein sequences and solved structures. In the genomic era, with the application of high-throughput DNA and protein sequencing technologies, the number of protein sequences has increased exponentially, at a pace that exceeds the pace at which protein structures are solved experimentally. Currently, only about 1.5% of protein sequences (about 40 000 out of 2.5 million known sequences available) have solved structures and the gap between proteins with known structures and with unknown structures is still increasing.

In spite of progress in robotics and other areas, experimental determination of a protein structure can still be expensive, labor intensive, time consuming, and not always possible. Some of the hardest challenges involve large quaternary complexes or particular classes of proteins, such as membrane proteins which are associated with a complex lipid bilayer environment. These proteins are particularly difficult to crystallize. Although membrane proteins are extremely important for biology and medicine, only a few dozen membrane protein structures are available in the PDB. Thus, in the remainder of this paper we focus almost exclusively on globular, nonmembrane proteins that are typically found in the cytoplasm or the nucleus of the cell, or that are secreted by the cell.

Protein structure prediction software is becoming an important proteomic tool for understanding phenomena in modern molecular and cell biology [18] and has important applications in biotechnology and medicine [19]. Here, we look at protein structure prediction at multiple levels, from 1-D to 4-D [20] and focus on the contributions made by machine learning approaches [21]. The 1-D prediction focuses on predicting structural features such as secondary structure [22]–[25] and relative solvent accessibility [26], [27] of each residue along the primary 1-D protein sequence (Fig. 2). The 2-D prediction focuses on predicting the spatial relationship between residues, such as distance and contact map prediction [28], [29] and disulfide bond prediction [30]–[33] (Fig. 3). One essential characteristic of these 2-D representations is that they are independent of any rotations and translations of the protein, therefore independent of any frame of coordinates, which appear only in the 3-D level. The 3-D prediction focuses on predicting the coordinates for all the residues or atoms of a protein in a 3-D space. Although the ultimate goal is to predict 3-D structure, 1-D and 2-D predictions are often used as input for 3-D coordinate predictors; furthermore, 1-D and 2-D predictions are also of intrinsic interest (Fig. 4). Finally, 4-D prediction focuses on the prediction of the structure of protein complexes comprised of several folded protein chains (Fig. 5).

The 1-D, 2-D, and 3-D protein structure prediction methods are routinely evaluated in the Critical Assessment of Techniques for the Protein Structure Prediction (CASP) [34] experiment—a community-wide experiment for blind protein structure prediction held every two years since 1994. The 4-D prediction methods are currently evaluated in the Critical Assessment of Techniques for Protein Interaction (CAPRI) [35]—a community-wide experiment for protein interaction. The assessment results are published in the supplemental issues of the journal *Proteins*.

To date, the most successful structure prediction methods have been knowledge based. Knowledge-based methods involve learning or extracting knowledge from existing solved protein structures and generalizing the gained knowledge to new proteins whose structures are unknown. Machine learning methods [21] that can automatically extract knowledge from the PDB are an important class of tools and have been widely used in all aspects of protein structure prediction. Here, we review the development and application of machine learning methods in 1-D, 2-D, 3-D, and 4-D structure prediction.

We focus primarily on unsupervised clustering methods and three supervised machine learning methods including hidden Markov models (HMMs) [21], [36], [37], neural networks [21], [38], and support vector machines [39] for 1-D, 2-D, 3-D, and 4-D structure prediction problems. We emphasize their applications to the problem of predicting the structure of globular proteins, which are the most abundant proteins—roughly 75% of a typical proteome—and for which several prediction methods have been developed. We also briefly review some applications of these methods to the prediction of the structure of membrane proteins, although far less training data is available for this class of proteins.

II. Machine Learning Methods for 1-D Structure Prediction

Many protein structural feature predictions are 1-D prediction problems, including, for example, secondary structure prediction, solvent accessibility prediction, disordered region predic-

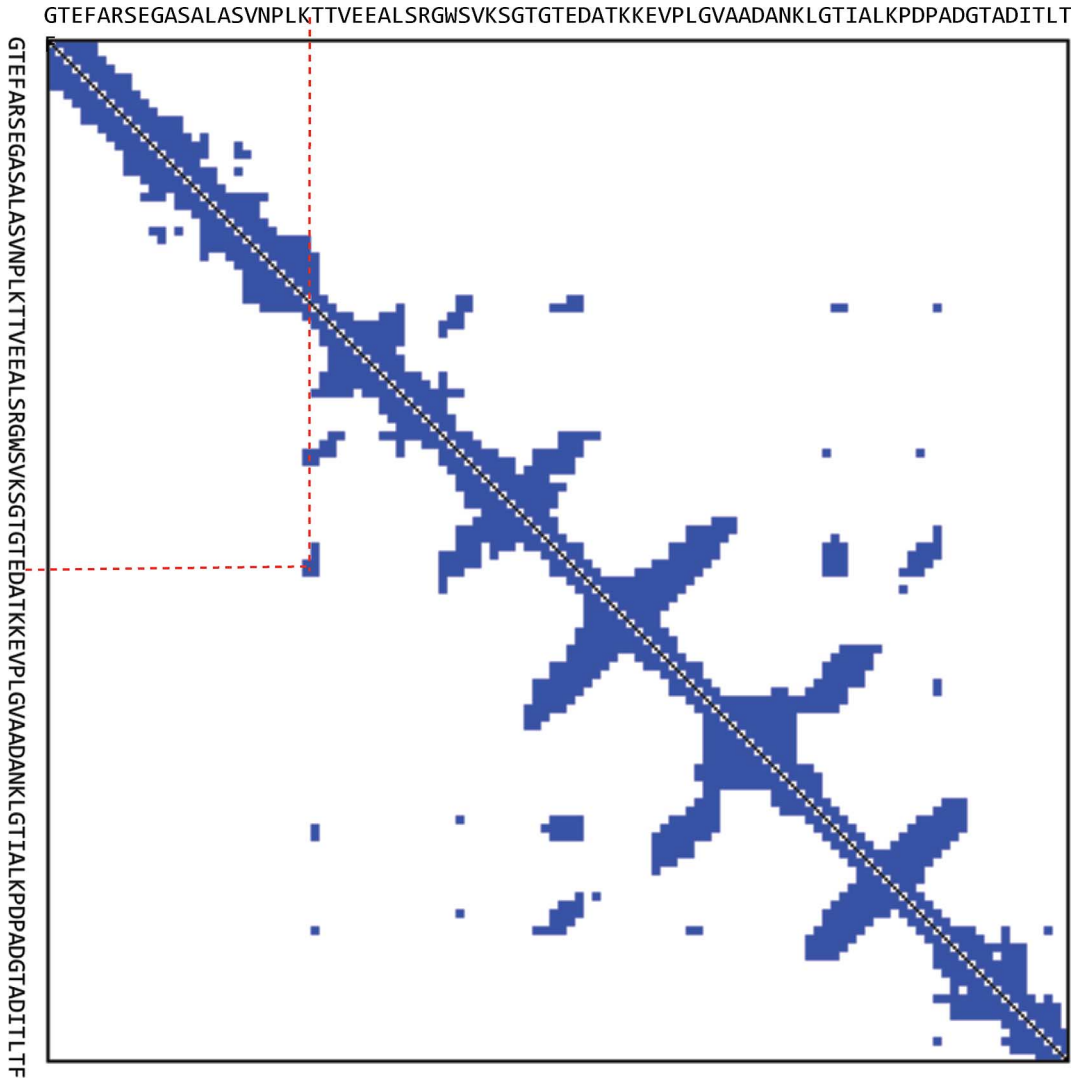

Fig. 3. Two-dimensional protein structure prediction. Example depicts a predicted 2-D contact map with an 8 Angstrom cutoff. The protein sequence is aligned along the sides of the contact map both horizontally and vertically. Each dot represents a predicted contact, i.e., a residue pair whose spatial distance is below 8 Angstroms. For instance, the red dotted lines mark a predicted contact associated with the pair (D, T).

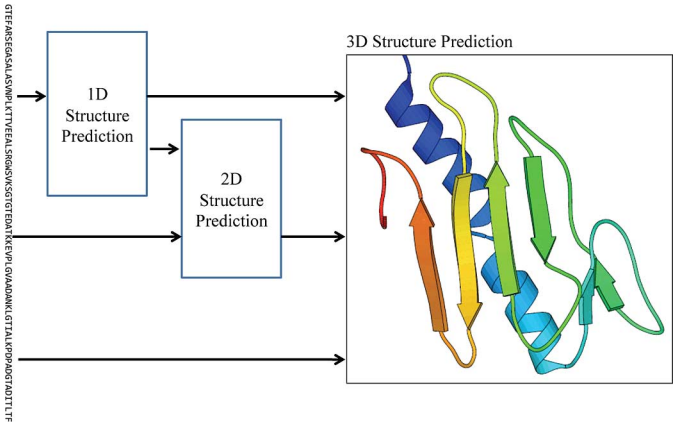

Fig. 4. Three-dimensional protein structure prediction. Three-dimensional structure predictors often combine information from the primary sequence and the predicted 1-D and 2-D structures to produce 3-D structure predictions.

tion, binding site prediction, functional site prediction, protein domain boundary prediction, and transmembrane helix prediction [22], [23], [33], [40]–[45], [96].

The input for 1-D prediction problems is a protein primary sequence and the output is a sequence of predicted features for each amino acid in the sequence. The learning goal is to map the input sequence of amino acids to the output sequence of features. The 1-D structure prediction problem is often viewed as a classification problem for each individual amino acid in the protein sequence. Historically, protein secondary structure prediction has been the most studied 1-D problem and has had a fundamental impact on the development of protein structure prediction methods [22], [23], [47]–[49]. Here, we will mainly focus on machine learning methods for secondary structure prediction of globular proteins. Similar techniques have also been applied to other 1-D prediction problems.

Early secondary structure prediction methods [47] were based on extracting the statistical correlations between a window of consecutive amino acids in a protein sequence and the secondary structure classification of the amino acid in the center of the window. Simple correlation methods capture a certain amount of information and can reach an accuracy of about 50%, well above chance levels. With the development

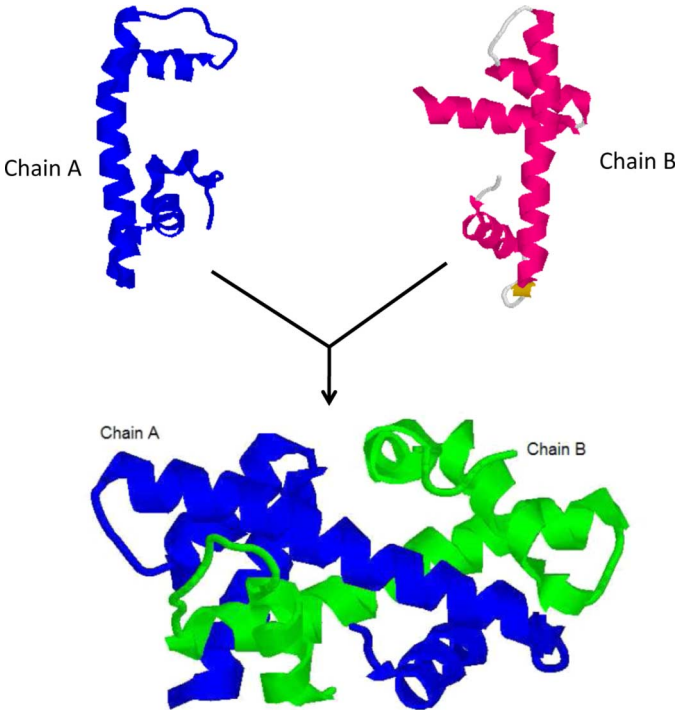

Fig. 5. Four-dimensional protein structure prediction. Four-dimensional prediction derived by docking individual protein chains to create a protein complex.

of more powerful pattern recognition and nonlinear function fitting methods, new approaches have been used to predict protein secondary structures. In the 1980s, feedforward neural networks were first applied to secondary structure prediction and significantly improved prediction accuracy to a level in the 60% to 70% range [48]. This was probably the first time a large-scale machine learning method was successfully applied to a difficult problem in bioinformatics. A third important breakthrough occurred with the realization that higher accuracy could be achieved by using a richer input derived from a multiple alignment of a sequence to its homologs. This is due to the fact that protein secondary structure is more conserved than protein primary sequence—i.e., protein sequences in the same protein family evolving from the same ancestor have different amino acid sequences but often maintain the same secondary structure [50], [51]. Rost and Sander [22], [23] were the first to combine neural networks with multiple sequence alignments to improve secondary structure prediction accuracy to about 70%–74%. In this approach, instead of encoding each amino acid with a sparse binary vector of length 20 containing a single 1-bit located at a different position for each different amino acid, the empirical probabilities (i.e., normalized frequencies) of the 20 amino acids appearing in the corresponding column of the multiple sequence alignment are used. The positional frequency vector, called the profile of the family at the corresponding position, captures evolutionary information related to the structural properties of the protein family. Profiles are relatively easy to create and allow one to leverage information contained in the sequence databases (e.g., SWISSPROT [53]) that are much larger than the PDB. Profiles are now used in virtually all knowledge-based protein structure prediction methods and have been further refined. For instance, PSI-PRED [24] uses PSI-BLAST [54] to derive new profiles based on position specific scoring matrices to further improve secondary structure prediction.

New algorithmic developments [49], [27] inspired by the theory of probabilistic graphical models [21] have led to more sophisticated recursive neural network architectures to try to improve prediction accuracy by incorporating information that extends beyond the fixed-size window input of traditional feedforward neural networks. Large ensembles of hundreds of neural networks have also been used [55]. The new technologies available along with the increase of protein sequence databases used to build profiles have improved secondary structure prediction accuracy to about 78%–80%. Moreover, hybrid methods [45], [57] that combine neural network approaches with homology searches have been developed to improve secondary structure prediction. Homologous proteins are proteins that are derived from the same evolutionary ancestor and therefore tend to share structural and functional characteristics. A protein that is strongly homologous to another protein with known structures in the PDB [15] will likely share a similar structure. In addition to neural networks, support vector machines (SVMs) are also another set of statistical machine learning methods used to predict protein secondary structures and other 1-D features of globular proteins with good accuracy [58].

Machine learning methods are also frequently used to predict 1-D feature of membrane proteins. For instance, neural networks as well as HMMs have been used to identify membrane proteins and predict their topology, which include predicting the location of their alpha-helical or beta-strand regions and the intracellular or extracellular localization of the loop regions [59], [96].

While 1-D prediction methods have made good progress over the past three decades, there is still room for some improvement in both the accuracy and scope of these methods. For instance, secondary structure prediction accuracy is still at least 8% below the predicted limit of 88% [60]. The prediction of protein domain boundaries [33], [40]–[42] and disordered regions [43]–[45] are still at an early stage of development, while already showing promising results. Some improvements may come from algorithmic improvements, for instance using ensemble and meta-learning techniques such as bagging and boosting [62] to combine classifiers to improve accuracy. Other improvements may require exploiting new sources of biological information. For instance, gene structure information, such as alternative splicing sites, may be used to improve domain boundary prediction [42].

III. Machine Learning Methods for 2-D Structure Prediction

The classic 2-D structure prediction problem is the prediction of protein contact maps [28], [63], [64]. A protein contact map is a matrix M, where $\mathrm{M}[\mathrm{i}, \mathrm{j}]$ is either one or zero, depending on whether the Euclidean distance between the two amino acids at linear positions i and j is above a specified distance threshold (e.g., 8 Angstroms) or not. Distances can be measured, for instance, between corresponding backbone carbon atoms. A coarser contact map can be derived in a similar way by considering secondary structure elements. Finer contact maps

can be derived by considering all the atoms of each amino acid. As previously mentioned, contact map representations are particularly interesting due to their invariance with respect to rotations and translations. Given an accurate contact map, several algorithms can be used to reconstruct the corresponding protein 3-D structure [65]–[67]. Since a contact map is essentially another representation of a protein 3-D structure, the difficulty of predicting a contact map is more or less equivalent to the difficulty of predicting the corresponding 3-D structure. Contact maps can also be used to try to infer protein folding rates [68], [69].

Several machine learning methods, including neural networks [28], [70]–[72], self-organizing maps [73], and support vector machines [74] have been applied to contact map prediction. Standard feedforward neural networks and support vector machines approaches use two windows around two target amino acids, i and j, to predict if they are in contact or not. This can be viewed as a binary classification problem. Each position in a window is usually a vector consisting of 20 numbers corresponding to the 20 profile probabilities, as in the 1-D prediction problem. Additional useful 1-D information that can be leveraged includes the predicted secondary structure or relative solvent accessibility of each amino acid. As in 1-D prediction, methods based on local windows approaches cannot take into account the effect of amino acids outside of the window. To overcome this problem, a 2-D-recursive neural network architecture [29] that in principle can use the entire sequence to derive each prediction was designed to improve contact map prediction. In the latest Critical Assessment of Techniques for protein structure prediction (CASP) [34], three methods using standard neural networks [72], 2-D recursive neural networks [45], and support vector machines [74] achieved the best results [75].

Despite progress made in the last several years, contact map prediction remains largely an unsolved problem. The current precision and recall of medium and long-range contact predictions is around 28% [74]. Although this number is quite low, its accuracy is better than the accuracy of contacts generated by other *ab initio* 3-D structure prediction methods. Predicted contact maps are likely to provide some help in 3-D structure prediction because even a small fraction of correctly predicted long-range contacts can effectively help build a protein topology [76].

In addition to the prediction of general residue-residue contact maps, special attention has been paid to more specific contact predictions: beta-strand pairing prediction [77] and disulfide bond prediction [33], [78], [79]. Disulfide bonds are covalent bonds that can form between cysteine residues. These disulfide bonds play a crucial role in stabilizing proteins, particularly small proteins. Disulfide bond prediction involves predicting if a disulfide bond exists between any two cysteine residues in a protein. Both neural networks and support vector machines have been used to predict disulfide bonds. The average precision and recall performance measures are slightly above 50%. Likewise, one can try to predict if two amino acids in two different beta-strands are paired or not in the same beta sheet. Usually, two paired beta-residues form hydrogen bonds with each other or their neighbors and contribute to the stabilization of the corresponding beta-sheet. In part because of the requirements imposed by the hydrogen bonding constraints, the accuracy of amino acid pairing in beta-sheets is above 41%, higher than the accuracy for generic contacts in contact maps. As with other 2-D prediction problems, feedforward and recursive neural networks have been used to predict beta-sheet pairings. Currently, the most successful method is a 2-D recursive neural network approach which takes a grid of beta-residues as inputs [77] and, together with graph matching algorithms, predicts pairings at the residue, strand, and sheet levels.

In addition to 2-D prediction for globular proteins, these techniques have recently been used to predict contacts in transmembrane beta-barrel proteins. Prediction of transmembrane beta-barrel proteins have been used to reconstruct 3-D structures with reasonable accuracy [59].

To use 2-D prediction more effectively as input features for 3-D structure prediction, one important task is to further improve 2-D prediction accuracy. As for 1-D predictions, progress may come from improvements in machine learning methods or, perhaps more effectively, from incorporating more informative features in the inputs. For instance, recently mutual information has been shown to be a useful feature for 2-D prediction [72], [74]. On the reconstruction side, several optimization algorithms exist to try to reconstruct 3-D structures from contact maps by using Monte Carlo methods [66], [82] and incorporating experimentally determined contacts or contacts extracted from template structures into protein structure prediction [83]–[85] or protein structure determination by NMR methods. However, these methods cannot reliably reconstruct 3-D structures from very noisy contact maps that were predicted from primary sequence information alone [66], [82]. Thus, of parallel importance is the development of more robust 3-D reconstruction algorithms that can tolerate the noise contained in predicted contact maps.

IV. Machine Learning Methods for 3-D Structure Prediction

Machine learning methods have been used in several aspects of protein 3-D structure prediction such as fold recognition [33], [86], [87], model generation [89], and model evaluation [90], [91].

Fold recognition aims to identify a protein, with known structure, that is presumably similar to the unknown structure of a query protein. Identification of structural homologs is an essential step for the most successful template-based 3-D structure prediction approaches. Neural networks were first used for this task in combination with threading [86]. More recently, a general machine learning framework has been proposed to improve both the sensitivity and specificity of fold recognition based on pairwise similarity features between query and template proteins [88]. Although the current implementation of the framework uses support vector machines to identify folds, it can be extended to any other supervised learning method.

In addition to classification methods, HMMs are among the most important techniques for protein fold recognition. Earlier HMM approaches, such as SAM [92] and HMMer [93], built an HMM for a query with its homologous sequences and then used this HMM to score sequences with known structures in the PDB

using the Viterbi algorithm, an instance of dynamic programming methods. This can be viewed as a form of profile-sequence alignment. More recently, profile–profile methods have been shown to significantly improve the sensitivity of fold recognition over profile–sequence, or sequence–sequence, methods [94]. In the HMM version of profile–profile methods, the HMM for the query is aligned with the prebuilt HMMs of the template library. This form of profile–profile alignment is also computed using standard dynamic programming methods.

Optimization techniques, such as conjugate gradient descent and Monte Carlo methods (e.g., simulated annealing) that are widely used in statistical machine learning methods are also essential techniques for 3-D protein structure generation and sampling. Conjugate gradient descent (a technique also used in neural network learning) is used to generate structures in the most widely used comparative modeling tool Modeller [95]. Lattice Monte Carlo sampling is used in both template-based and *ab initio* structure modeling [85] and the most widely used *ab initio* fragment assembly tool, Rosetta, uses simulated annealing sampling techniques [89].

In addition to model generation, machine learning methods are also widely used to evaluate and select protein models. Most *ab initio* structure prediction methods use clustering techniques to select models [96]. These methods first generate a large population of candidate models and then cluster them into several clusters based on the structure similarity between the models, using k-means clustering or some other similar clustering algorithm. Representative elements from each cluster, such as the centroids, are then proposed as possible 3-D structures. Usually, the centroid of the largest cluster is used as the most confident prediction, although occasionally the centroid of a different cluster can be even closer to the native structure. In addition to clustering, supervised learning techniques have been used to directly assess the quality of a protein model. Neural networks have been used to estimate the root mean square distance (RMSD) between a model and the native structure [90]. Support vector machines have been used to rank protein models [91]. One main challenge of model selection is that current methods cannot consistently select the best model with lowest RMSD. For model quality evaluation, the correlation between predicted scores and real quality scores for hard targets (poor models) is still low [97], i.e., some poor models may receive good predicted scores. In addition, a statistical confidence score should be assigned to the predicted quality scores for better model usage and interpretation. It is likely that additional machine learning methods will have to be developed to better deal with these problems.

V. Machine Learning Methods for 4-D Structure Prediction

The aim of 4-D structure prediction is to predict the structure of a protein complex consisting of two or more protein chains, also known as protein docking [98]–[106]. Like 3-D structure prediction, 4-D structure prediction is often reduced to a problem of conformation sampling with the use of energy functions [107]–[110].

Assuming the 3-D structures of each protein subunit are known, some docking methods use 3-D grid Fourier transformation methods [111] to dock protein subunits together. More recently, RosettaDock uses the same simulated annealing technique as Rosetta for 3-D, with some adjustments to the 4-D problem [106]. More broadly, several ongoing efforts aim to adapt 3-D methods to 4-D problems. For instance, clustering methods have been adapted to cluster docking conformations and to select centroids of clusters to generate final predictions [112].

Four-dimensional prediction is closely related to 1-D, 2-D, and 3-D prediction. For instance, if the protein interaction interfaces (sites) can be accurately predicted by 1-D predictors [113], the conformation search space for the protein docking phase can be drastically reduced. Since one of the major bottlenecks of 4-D prediction is the size of the conformation space to be sampled, which is even larger than in the 3-D case, improving interface prediction is an essential step to address this bottleneck. Currently, neural networks, HMMs and support vector machine methods have been used to predict interface sites [114]. Most of these methods use some features extracted from the 3-D structures of the protein subunits. Since in most practical cases the 3-D structures themselves are currently not available, it may be worthwhile to further develop methods to predict interactions from protein sequences alone.

The other major bottleneck in protein docking comes from induced conformational changes, which introduce an additional layer of complexity that is not well handled by current methods [107]. Most current docking methods assume that the structures of the subunits are subjected to little or no changes during docking. However, upon protein binding, individual proteins may undergo substantial or even large-scale conformational changes, which cannot be handled by current docking methods. Developing machine learning methods to identify regions, such as flexible hinges, that facilitate large-scale movement may be of some help in predicting the overall structure of these protein complexes, although the amount of available training data for this problem may not be as abundant as one would like.

Finally, as in the case of 3-D structure prediction, machine learning methods may help in developing better methods for assessing the quality of 4-D models and predict their quality and confidence levels.

VI. Conclusion

Machine learning methods have played, and continue to play, an important role in 1-D-4-D protein structure predictions, as well as in many other related problems. For example, machine learning methods are being used to predict protein solubility [115], protein stability [116], protein signal peptides [117], [118], protein cellular localization [117], protein post-translation modification sites, such as phosphorilation sites [119], and protein epitopes [120]–[123]. Here, we have tried to give a selected and nonexhaustive overview of some of the applications of machine learning methods to protein structure prediction problems.

A common question often asked by students is which machine learning method is "better" or more suitable for a given problem? In short, should I use a neural network, an HMM, an SVM, or something else? In our opinion, it turns out that this question is not as fundamental as it may seem. While a given machine learning approach may be easier to implement for a given problem, or more suited to a particular data format, to

tackle difficult problems what matters in the end is the expertise a scientist has in a particular machine learning technology. What can be obtained with a general-purpose machine learning method can be achieved using another general-purpose machine learning method, provided the learning architecture and algorithms are properly crafted.

In the foreseeable future, machine learning methods will continue to play a role in protein structure prediction and its multiple facets. The growth in the size of the available training sets coupled with the gap between the number of sequences and the number of solved structures remain powerful motivators for further developments. Furthermore, in many cases machine learning methods are relatively fast compared to other methods. Machine learning methods spend most of their time in the learning phase, which can be done offline. In "production" mode, a pretrained feedforward neural network, for instance, can produce predictions rather fast. Both accuracy and speed considerations are likely to remain important as genomic, proteomic, and protein engineering projects continue to generate great challenges and opportunities in this area.

REFERENCES

[1] F. Sanger and E. O. Thompson, "The amino-acid sequence in the glycyl chain of insulin.1. The identification of lower peptides from partial hydrolysates," *J. Biochem.*, vol. 53, no. 3, pp. 353–366, 1953a.

[2] F. Sanger and E. O. Thompson, "The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates," *J. Biochem.*, vol. 53, no. 3, pp. 366–374, 1953b.

[3] L. Pauling and R. B. Corey, "The pleated sheet, a new layer configuration of the polypeptide chain," *Proc. Nat. Acad. Sci.*, vol. 37, pp. 251–256, 1951.

[4] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: Two hydrogenbonded helical configurations of the polypeptide chain," *Proc. Nat. Acad. Sci.*, vol. 37, pp. 205–211, 1951.

[5] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. J. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore, "Structure of myoglobin: A three-dimensional Fourier synthesis at 2_a resolution," *Nature*, vol. 185, pp. 422–427, 1960.

[6] M. F. Perutz, M. G. Rossmann, A. F. Cullis, G. Muirhead, G. Will, and A. T. North, "Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 Angstrom resolution, obtained by x-ray analysis," *Nature*, vol. 185, pp. 416–422, 1960.

[7] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 31, pp. 7134–7155, 1990.

[8] R. A. Laskowski, J. D. Watson, and J. M. Thornton, "From protein structure to biochemical function?," *J. Struct. Funct. Genomics*, vol. 4, pp. 167–177, 2003.

[9] A. Travers, "DNA conformation and protein binding," *Ann. Rev. Biochem.*, vol. 58, pp. 427–452, 1989.

[10] P. J. Bjorkman and P. Parham, "Structure, function and diversity of class I major histocompatibility complex molecules," *Ann. Rev. Biochem.*, vol. 59, pp. 253–288, 1990.

[11] L. Bragg, *The Development of X-Ray Analysis*. London, U.K.: G. Bell, 1975.

[12] T. L. Blundell and L. H. Johnson, *Protein Crystallography*. New York: Academic, 1976.

[13] K. Wuthrich, *NMR of Proteins and Nucleic Acids*. New York: Wiley, 1986.

[14] E. N. Baldwin, I. T. Weber, R. S. Charles, J. Xuan, E. Appella, M. Yamada, K. Matsushima, B. F. P. Edwards, G. M. Clore, A. M. Gronenborn, and A. Wlodawar, "Crystal structure of interleukin 8: Symbiosis of NMR and crystallography," *Proc. Nat. Acad. Sci.*, vol. 88, pp. 502–506, 1991.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucl. Acids Res.*, vol. 28, pp. 235–242, 2000.

[16] J. M. Chandonia and S. E. Brenner, "The impact of structural genomics: Expectations and outcomes," *Science*, vol. 311, pp. 347–351, 2006.

[17] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, pp. 223–230, 1973.

[18] D. Petrey and B. Honig, "Protein structure prediction: Inroads to biology," *Mol. Cell.*, vol. 20, pp. 811–819, 2005.

[19] M. Jacobson and A. Sali, "Comparative protein structure modeling and its applications to drug discovery," in *Annual Reports in Medical Chemistry*, J. Overington, Ed. London, U.K.: Academic, 2004, pp. 259–276.

[20] B. Rost, J. Liu, D. Przybylski, R. Nair, K. O. Wrzeszczynski, H. Bigelow, and Y. Ofran, "Prediction of protein structure through evolution," in *Handbook of Chemoinformatics – From Data to Knowledge*, J. Gasteiger and T. Engel, Eds. New York: Wiley, 2003, pp. 1789–1811.

[21] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, 2nd ed. Cambridge, MA: MIT Press, 2001.

[22] B. Rost and C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," *Proc. Nat. Acad. Sci.*, vol. 90, no. 16, pp. 7558–7562, 1993a.

[23] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *J. Mol. Bio.*, vol. 232, no. 2, pp. 584–599, 1993b.

[24] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Bio.*, vol. 292, pp. 195–202, 1999b.

[25] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins*, vol. 47, pp. 228–235, 2002a.

[26] B. Rost and C. Sander, "Conservation and prediction of solvent accessibility in protein families," *Proteins*, vol. 20, no. 3, pp. 216–226, 1994.

[27] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, "Prediction of coordination number and relative solvent accessibility in proteins," *Proteins*, vol. 47, pp. 142–153, 2002b.

[28] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, "Prediction of contact maps with neural networks and correlated mutations," *Prot. Eng.*, vol. 13, pp. 835–843, 2001.

[29] G. Pollastri and P. Baldi, "Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners," *Bioinfo.*, vol. 18, no. Suppl 1, pp. S62–S70, 2002.

[30] P. Fariselli and R. Casadio, "Prediction of disulfide connectivity in proteins," *Bioinfo.*, vol. 17, pp. 957–964, 2004.

[31] A. Vullo and P. Frasconi, "A recursive connectionist approach for predicting disulfide connectivity in proteins," in *Proc. 18th Annu. ACM Symp. Applied Computing*, 2003, pp. 67–71.

[32] P. Baldi, J. Cheng, and A. Vullo, "Large-scale prediction of disulphide bond connectivity," in *Advances in Neural Information Processing Systems*, L. Bottou, L. Saul, and Y. Weiss, Eds. Cambridge, MA: MIT Press, 2005, vol. 17, NIPS04 Conf., pp. 97–104.

[33] J. Cheng, H. Saigo, and P. Baldi, "Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching," *Proteins: Structure, Function, Bioinformatics*, vol. 62, no. 3, pp. 617–629, 2006b.

[34] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, and A. Tramontano, "Critical assessment of methods of protein structure prediction-Round VII," *Proteins*, vol. 29, pp. 179–187, 2007.

[35] S. J. Wodak, "From the Mediterranean coast to the shores of Lake Ontario: CAPRI's premiere on the American continent," *Proteins*, vol. 69, pp. 687–698, 2007.

[36] P. Baldi, Y. Chauvin, T. Hunkapillar, and M. McClure, "Hidden Markov models of biological primary sequence information," *Proc. Nat. Acad. Sci.*, vol. 91, no. 3, pp. 1059–1063, 1994.

[37] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *J. Mol. Biol.*, vol. 235, pp. 1501–1531, 1994.

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating error," *Nature*, vol. 323, pp. 533–536, 1986.

[39] V. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag, 1995.

[40] K. Bryson, D. Cozzetto, and D. T. Jones, "Computer-assisted protein domain boundary prediction using the DomPred server," *Curr Protein Pept Sci.*, vol. 8, pp. 181–188, 2007.

[41] M. Tress, J. Cheng, P. Baldi, K. Joo, J. Lee, J. H. Seo, J. Lee, D. Baker, D. Chivian, D. Kim, A. Valencia, and I. Ezkurdia, "Assessment of predictions submitted for the CASP7 domain prediction category," *Proteins: Structure, Function and Bioinformatics*, vol. 68, no. S8, pp. 137–151, 2007.

[42] J. Cheng, "DOMAC: An accurate, hybrid protein domain prediction server," *Nucleic Acids Res.*, vol. 35, pp. w354–w356, 2007.

[43] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker, "Exploiting heterogeneous sequence properties improves prediction of protein disorder," *Proteins*, vol. 61, no. Suppl 1, pp. 176–182, 2005.

[44] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones, "The DISOPRED server for the prediction of protein disorder," *Bioinfo.*, vol. 20, pp. 2138–2139, 2004.

[45] J. Cheng, M. J. Sweredoski, and P. Baldi, "Accurate prediction of protein disordered regions by mining protein structure data," *Data Mining Knowledge Discovery*, vol. 11, pp. 213–222, 2005.

[46] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes," *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, 2001.

[47] P. Y. Chou and G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence," *Adv. Enzymol.*, vol. 47, pp. 45–148, 1978.

[48] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *J. Mol. Biol.*, vol. 202, pp. 265–884, 1988.

[49] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future in protein secondary structure prediction," *Bioinformatics*, vol. 15, no. 11, pp. 937–946, 1999.

[50] I. P. Crawford, T. Niermann, and K. Kirchner, "Prediction of secondary structure by evolutionary comparison: Application to the a subunit of tryptophan synthase," *Proteins*, vol. 2, pp. 118–129, 1987.

[51] G. J. Barton, R. H. Newman, P. S. Freemont, and M. J. Crumpton, "Amino acid sequence analysis of the annexin supergene family of proteins," *Eur. J. Biochem.*, vol. 198, pp. 749–760, 1991.

[52] B. Rost and C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," *Proc. Nat. Acad. Sci.*, vol. 90, pp. 7558–7562, 1993.

[53] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh, "The universal protein resource (UniProt)," *Nucleic Acids Res.*, vol. 33, pp. D154–159, 2005.

[54] S. F. Altschul, T. L. Madden, A. A. Scha_er, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nuc. As. Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[55] G. Pollastri and A. McLysaght, "Porter: A new, accurate server for protein secondary structure prediction," *Bioinfo.*, vol. 21, no. 8, pp. 1719–20, 2005.

[56] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, "SCRATCH: A protein structure and structural feature prediction server," *Nuc. As. Res.*, vol. 33, pp. 72–76, 2005.

[57] R. Bondugula and D. Xu, "MUPRED: A tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction," *Proteins*, vol. 66, no. 3, pp. 664–670, 2007.

[58] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Secondary structure prediction using support vector machines," *Bioinfo.*, vol. 19, pp. 1650–1655, 2003.

[59] Randall, J. Cheng, M. Sweredoski, and P. Baldi, "TMBpro: Secondary structure, beta-contact, and tertiary structure prediction of transmembrane beta-barrel proteins," *Bioinfo.*, vol. 24, pp. 513–520, 2008.

[60] B. Rost, "Rising accuracy of protein secondary structure prediction," in *Protein Structure Determination, Analysis, and Modeling for Drug Discovery*, D. Chasman, Ed. New York: Marcel Dekker, 2003, pp. 207–249.

[61] J. Cheng, M. Sweredoski, and P. Baldi, "DOMpro: Protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks," *Data Mining Knowledge Discovery*, vol. 13, pp. 1–10, 2006.

[62] Y. Freund, "Boosting a weak learning algorithm by majority," in *Proc. Third Annu. Workshop Computational Learning Theory*, 1990.

[63] O. Olmea and A. Valencia, "Improving contact predictions by the combination of correlated mutations and other sources of sequence information," *Fold Des*, vol. 2, pp. s25–s32, 1997.

[64] P. Baldi and G. Pollastri, "A machine learning strategy for protein analysis," *IEEE Intelligent Systems, Special Issue Intelligent Systems in Biology*, vol. 17, no. 2, pp. 28–35, Feb. 2002.

[65] A. Aszodi, M. Gradwell, and W. Taylor, "Global fold determination from a small number of distance restraints," *J. Mol. Biol.*, vol. 251, pp. 308–326, 1995.

[66] M. Vendruscolo, E. Kussell, and E. Domany, "Recovery of protein structure from contact maps," *Folding Design*, vol. 2, pp. 295–306, 1997.

[67] J. Skolnick, A. Kolinski, and A. Ortiz, "MONSSTER: A method for folding globular proteins with a small number of distance restraints," *J Mol. Biol.*, vol. 265, pp. 217–241, 1997.

[68] K. Plaxco, K. Simons, and D. Baker, "Contact order, transition state placement and the refolding rates of single domain proteins," *J. Mol. Biol.*, vol. 277, pp. 985–994, 1998.

[69] M. Punta and B. Rost, "Protein folding rates estimated from contact predictions," *J. Mol. Biol.*, pp. 507–512, 2005a.

[70] P. Baldi and G. Pollastri, "The principled design of large-scale recursive neural network architectures—DAG-RNNs and the protein structure prediction problem," *J. Machine Learning Res.*, vol. 4, pp. 575–602, 2003.

[71] M. Punta and B. Rost, "PROFcon: Novel prediction of long-range contacts," *Bioinfo.*, vol. 21, pp. 2960–2968, 2005b.

[72] G. Shackelford and K. Karplus, "Contact prediction using mutual information and neural nets," *Proteins*, vol. 69, pp. 159–164, 2007.

[73] R. MacCallum, "Striped sheets and protein contact prediction," *Bioinfo.*, vol. 20, no. Supplement 1, pp. i224–i231, 2004.

[74] J. Cheng and P. Baldi, "Improved residue contact prediction using support vector machines and a large feature set," *BMC Bioinformatics*, vol. 8, p. 113, 2007.

[75] J. M. G. Izarzugaza, O. Graña, M. L. Tress, A. Valencia, and N. D. Clarke, "Assessment of intramolecular contact predictions for CASP7," *Proteins*, vol. 69, pp. 152–158.

[76] S. Wu and Y. Zhang, "A comprehensive assessment of sequence-based and template-based methods for protein contact prediction," *Bioinfo.*, 2008, to be published.

[77] J. Cheng and P. Baldi, "Three-stage prediction of protein beta-sheets by neural networks, alignments, and graph algorithms," *Bioinfo.*, vol. 21, pp. i75–i84, 2005.

[78] P. Fariselli, P. Riccobelli, and R. Casadio, "Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins," *Proteins*, vol. 36, pp. 340–346, 1999.

[79] A. Vullo and P. Frasconi, "Disulfide connectivity prediction using recursive neural networks and evolutionary information," *Bioinfo.*, vol. 20, pp. 653–659, 2004.

[80] J. Cheng, H. Saigo, and P. Baldi, "Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching," *Proteins: Structure, Function, Bioinformatics*, vol. 62, no. 3, pp. 617–629, 2006b.

[81] A. Z. Randall, J. Cheng, M. Sweredoski, and P. Baldi, "TMBpro: Secondary structure, beta-contact, and tertiary structure prediction of transmembrane beta-barrel proteins," *Bioinfo.*, vol. 24, pp. 513–520, 2008.

[82] M. Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli, and R. Casadio, "FT-COMAR: Fault tolerant three-dimensional structure reconstruction from protein contact maps," *Bioinfo.*, vol. 24, pp. 1313–1315, 2008.

[83] C. A. Rohl and D. Baker, "De novo determination of protein backbone structure from residual dipolar couplings using Rosetta," *J. Amer. Chemical Soc.*, vol. 124, pp. 2723–2729, 2004.

[84] P. M. Bowers, C. E. Strauss, and D. Baker, "De novo protein structure determination using sparse NMR data," *J. Biomol. NMR*, vol. 18, no. 4, pp. 311–318, 2000.

[85] Y. Zhang and J. Skolnick, "Automated structure prediction of weakly homologous proteins on a genomic scale," *Proc Nat. Acad. Sci.*, vol. 101, no. 20, pp. 7594–7599, 2004a.

[86] D. T. Jones, "GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences," *J. Mol. Biol.*, vol. 287, pp. 797–815, 1999a.

[87] D. Kim, D. Xu, J. Guo, K. Ellrott, and Y. Xu, "PROSPECT II: Protein structure prediction method for genome-scale applications," *Protein Eng.*, vol. 16, no. 9, pp. 641–650, 2003.

[88] J. Cheng and P. Baldi, "A machine learning information retrieval approach to protein fold recognition," *Bioinfo.*, vol. 22, no. 12, pp. 1456–1463, 2006.

[89] 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," *J. Mol. Biol.*, vol. 268, pp. 209–225, 1997.

[90] B. Wallner and A. Elofsson, "Prediction of global and local model quality in CASP7 using Pcons and ProQ," *Proteins*, vol. 69, pp. 184–193, 2007.

[91] J. Qiu, W. Sheffler, D. Baker, and W. S. Noble, "Ranking predicted protein structures with support vector regression," *Proteins*, vol. 71, pp. 1175–1182, 2007.

[92] K. Karplus, C. Barrett, and R. Hughey, "Hidden Markov models for detecting remote protein homologies," *Bioinfo.*, vol. 14, no. 10, pp. 846–856, 1998.

[93] S. R. Eddy, "Profile hidden Markov models," *Bioinfo.*, vol. 14, pp. 755–763, 1998.

[94] J. Soeding, "Protein homology detection by HMM-HMM comparison," *Bioinfo.*, vol. 21, pp. 951–960, 2005.

[95] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *J. Mol. Biol.*, vol. 234, pp. 779–815, 1993.

[96] Y. Zhang and J. Skolnick, "SPICKER: A clustering approach to identify near-native protein folds," *J. Comp. Chem.*, vol. 25, pp. 865–871, 2004b.

[97] D. Cozzetto, A. Kryshtafovych, M. Ceriani, and A. Tramontano, "Assessment of predictions in the model quality assessment category," *Proteins*, vol. 69, no. S8, pp. 175–183, 2007.

[98] P. Aloy, G. Moont, H. A. Gabb, E. Querol, F. X. Aviles, and M. J. E. Sternberg, "Modelling protein docking using shape complementarity, electrostatics and biochemical information," *Proteins*, vol. 33, pp. 535–549, 1998.

[99] A. J. Bordner and A. A. Gorin, "Protein docking using surface matching and supervised machine learning," *Proteins*, vol. 68, pp. 488–502, 2007.

[100] V. Chelliah, T. L. Blundell, and J. Fernandez-Recio, "Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment," *J. Mol. Biol.*, vol. 357, pp. 1669–1682, 2006.
[101] R. Chen, L. Li, and Z. Weng, "ZDOCK: An initial-stage protein docking algorithm," *Proteins*, vol. 52, pp. 80–87, 2003.
[102] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: An automated docking and discrimination method for the prediction of protein complexes," *Bioinfo.*, vol. 20, pp. 45–50, 2004.
[103] M. D. Daily, D. Masica, A. Sivasubramanian, S. Somarouthu, and J. J. Gray, "CAPRI rounds 3–5 reveal promising successes and future challenges for RosettaDock," *Proteins*, vol. 60, pp. 181–186, 2005.
[104] C. Dominguez, R. Boelens, and A. Bonvin, "HADDOCK: A protein-protein docking approach based on biochemical or biophysical information," *J. Amer. Chem. Soc.*, vol. 125, pp. 1731–1737, 2003.
[105] H. A. Gabb, R. M. Jackson, and M. J. E. Sternberg, "Modelling protein docking using shape complementarity, electrostatics, and biochemical information," *J. Mol. Biol.*, vol. 272, pp. 106–120, 1997.
[106] J. J. Gray, S. E. Moughan, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker, "Protein-protein docking with simultaneous optimization of rigid body displacement and side chain conformations," *J. Mol. Biol.*, vol. 331, pp. 281–299, 2003.
[107] S. J. Wodak and R. Mendez, "Prediction of protein-protein interactions: The CAPRI experiment, its evaluation and implications," *Curr. Opin. Struct. Biol.*, vol. 14, pp. 242–249, 2004.
[108] H. Lu, L. Lu, and J. Skolnick, "Development of unified statistical potentials describing protein-protein interactions," *Biophysical J.*, vol. 84, pp. 1895–1901, 2003.
[109] J. Mintseris, B. Pierce, K. Wiehe, R. Anderson, R. Chen, and Z. Weng, "Integrating statistical pair potentials into protein complex prediction," *Proteins*, vol. 69, pp. 511–520, 2007.
[110] G. Moont, H. A. Gabb, and M. J. Sternberg, "Use of pair potentials across protein interfaces in screening predicted docked complexes," *Proteins*, vol. 35, pp. 364–373, 1999.
[111] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser, "Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques," *Proc. Nat. Acad. Sci.*, vol. 89, pp. 2195–2199, 1992.
[112] S. Lorenzen and Y. Zhang, "Identification of near-native structures by clustering protein docking conformations," *Proteins*, vol. 68, pp. 187–194, 2007.
[113] H. X. Zhou and S. Qin, "Interaction-site prediction for protein complexes: A critical assessment," *Bioinfo.*, vol. 23, no. 17, pp. 2203–2209, 2007.
[114] H. X. Zhou and Y. Shan, "Prediction of protein interaction sites from sequence profile and residue neighbor list," *Proteins*, vol. 44, pp. 336–343, 2001.
[115] P. Smialowski, A. J. Martin-Galiano, A. Mikolajka, T. Girschick, T. A. Holak, and D. Frishman, "Protein solubility: Sequence based prediction and experimental verification," *Bioinformatics*, vol. 23, pp. 2536–2542, 2007.
[116] J. Cheng, A. Randall, and P. Baldi, "Prediction of protein stability changes for single site mutations using support vector machines," *Proteins*, vol. 62, no. 4, pp. 1125–1132, 2006c.
[117] O. Emanuelsson, S. Brunak, G. V. Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP, and related tools," *Nature Protocols*, vol. 2, pp. 953–971, 2007.
[118] J. D. Bendtsen, H. Nielsen, G. V. Heijne, and S. Brunak, "Improved prediction of signal peptides: SignalP 3.0," *J. Mol. Biol.*, vol. 340, pp. 783–795, 2004.
[119] N. Blom, S. Gammeltoft, and S. Brunak, "Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites," *J. Molecular Biol.*, vol. 294, pp. 1351–1362, 1999.
[120] P. H. Andersen, M. Nielsen, and O. Lund, "Prediction of residues in discontinuous B-cell epitopes using protein 3D structures," *Protein Sci.*, vol. 15, pp. 2558–2567, 2006.
[121] J. Larsen, O. Lund, and M. Nielsen, "Improved method for predicting linear B-cell epitopes," *Immunome Res.*, vol. 2, p. 2, 2006.
[122] J. Sweredoski and P. Baldi, "PEPITO: Improved discontinuous B-cell epitope prediction using multiple distance thresholds and half-sphere exposure," *Bioinformatics*, vol. 24, pp. 1459–1460, 2008a.
[123] J. Sweredoski and P. Baldi, *COBEpro: A Novel System for Predicting Continuous B-Cell Epitopes*, 2008, submitted for publication.

Jianlin Cheng received the Ph.D. degree from the University of California, Irvine, 2006.

He is an Assistant Professor of bioinformatics in the Computer Science Department, University of Missouri, Columbia (MU). He is affiliated with the MU Informatics Institute, the MU Interdisciplinary Plant Group, and the National Center for Soybean Biotechnology. His research is focused on bioinformatics, systems biology, and machine learning.

Allison N. Tegge (M'08) received the B.Sc. degree in animal science and the M.Sc. degree in bioinformatics, both from the University of Illinois, Urbana-Champaign. She is working toward the Ph.D. degree in bioinformatics at the University of Missouri, Columbia (MU).

She is a National Library of Medicine (NLM) Fellow. Her research interests include protein structure prediction and systems biology.

Pierre Baldi (M'88–SM'01) received the Ph.D. degree from the California Institute of Technology, in 1986.

He is the Chancellor's Professor in the School of Information and Computer Sciences and the Department of Biological Chemistry and the Director of the UCI Institute for Genomics and Bioinformatics at the University of California, Irvine. From 1986 to 1988, he was a Postdoctoral Fellow at the University of California, San Diego. From 1988 to 1995, he held faculty and member of the technical staff positions at the California Institute of Technology and at the Jet Propulsion Laboratory. He was CEO of a startup company from 1995 to 1999 and joined UCI in 1999. His research work is at the intersection of the computational and life sciences, in particular the application of AI/statistical/machine learning methods to problems in bio- and chemical informatics. He has published over 200 peer-reviewed research articles and four books: *Modeling the Internet and the We–Probabilistic Methods and Algorithms* (Wiley, 2003); *DNA Microarrays and Gene Regulation–From Experiments to Data Analysis and Modeling* (Cambridge University Press, 2002); *The Shattered Self–The End of Evolution*, (MIT Press, 2001); and *Bioinformatics: the Machine Learning Approach* (MIT Press, Second Edition, 2001).

Dr. Baldi is the recipient of a 1993 Lew Allen Award, a 1999 Laurel Wilkening Faculty Innovation Award, and a 2006 Microsoft Research Award and was elected an AAAI Fellow in 2007.