Sequence-Based Methods for Real Value Predictions of Protein Structure

Lukasz Kurgan[*,1], Krzysztof Cios[2,5], Hua Zhang[3,1], Tuo Zhang[3,1], Ke Chen[1], Shiyi Shen[3,4] and Jishou Ruan[3,4]

[1]*Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada;* [2]*Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA;* [3]*College of Mathematical Science and LPMC, Nankai University, Tianjin, PRC;* [4]*Chern Institute of Mathematics, Tianjin, PRC;* [5]*IITiS, PAN, Poland*

Abstract: Recent years observed a growing interest in computational methods that predict and characterize protein structure due to the increasing sequence-structure gap. This includes a spike in development of sequence-based in-silico methods that address prediction of several newly formulated real-value descriptors of protein structure. These descriptors include B-factor, backbone torsion angles, solvent accessibility, residue depth, contact number, residue-wise contact order, secondary structure content, and folding rates. Although they address different structural aspects, such as exposure to the solvent, spatial position and packing of the residues, their flexibility, amount of secondary structures in the protein, and folding time, the methods that are built to address them share similarities that could be exploited to improve future designs. To date, no comprehensive overview that summarizes and contrasts solutions developed for these tasks was published. To address this we compare different designs of real-value predictors based on information concerning input data encoding and prediction algorithms used. We also investigate evaluation standards, which include benchmark datasets, test criteria, and test procedures used in these predictive tasks. Finally, we summarize application areas and problems that use the above-mentioned predictions. We believe that the breath and number of these applications justify further development of more accurate and integrated real-value prediction methods.

Keywords: Real-value prediction, protein structure, solvent accessibility, residue depth, contact number, residue-wise contact order, secondary structure content, backbone torsion angles, B-factor, folding rate.

INTRODUCTION

The number of computational systems that address characterization and prediction of structural aspects of proteins is growing rapidly. Their development is motivated by the widening structure-sequence gap and the growing size of the databases, such as PDB [1], SWISS-PROT [2], SCOP [3], CATH [4], etc., which organize and provide access to experimentally-derived protein structures. The methods utilize information about known structures of roughly 50 thousand proteins (as of May 2008 50,900 protein structures were stored in the PDB) to find regularities, which are encapsulated in models that can be used to perform high-throughput prediction of structures of over 5 millions proteins for which only the sequence is known (as of May 2008 5.4 million non-redundant protein sequences are stored in the NCBI's RefSeq database [5]).

This review paper focuses on methods that use protein sequence to predict various aspects of structure of globular protein that are encoded using real values. Such descriptors quantify *local properties* of the amino acids such as their mobility in the crystal structure (B-factor), backbone torsion angle, exposure to the solvent (solvent accessibility, residue depth), number of neighboring residues (contact number), linear (along the protein sequence) distance to the neighboring residues (residue-wise contact order), and *global properties* that concern the entire protein such as the secondary structure content and the folding rate. We observe that although the abovementioned prediction methods address a variety of different objectives, they also share a number of similarities with respect to their designs and protocols used in their evaluation. We will summarize and contrast the sequence representations and prediction algorithms that are used in the above-mentioned works, as well as the datasets, quality indices, and test procedures used to evaluate their performance. Additionally, we will review applications that use these methods.

Although this survey concentrates on globular proteins, we emphasize that a large body of research is devoted to development of prediction methods that address membrane proteins, which account for approximately 30% of the known proteins. The representative targets in the context of the membrane proteins include prediction of membrane protein types [6-12], transmembrane helices [13, 14], and amenability of membrane proteins for cloning, expression, and solubilization [15].

BACKGROUND

This section defines real-value protein structure descriptors and summarizes methods that predict the defined descriptor values using protein sequence as the input. In general, any prediction method consists of two steps: 1) the protein sequence is converted into a feature-based representation, and 2) the feature values are fed into a prediction algorithm to produce the values of the descriptors.

B-factor

The B-factor (also called B-value, Debye–Waller factor, or temperature factor) is used to measure local flexibility

*Address correspondence to this author at the Department of Electrical and Computer Engineering, ECERF (9107 116 Street), University of Alberta, Edmonton, AB, Canada T6G 2V4, Canada; Fax: (780) 492-1811;
E-mail: lkurgan@ece.ualberta.ca

1574-8936/08 $55.00+.00

(mobility) of residues. B-factor values are reported from experimental atomic-resolution structures. They quantify the decrease of intensity in diffraction due to the dynamic disorder caused by the temperature-dependent vibration of the atoms and the static disorder related to orientation of the protein molecule. High values indicate higher mobility of residues in crystal structures. The B-factor of i^{th} residue is defined as:

$$B_i = 8\pi\langle u_i^2 \rangle$$

where $\langle u_i^2 \rangle$ is the unidirectional mean-square displacement averaged over the lattice [16]. The B-factor values of C_α atoms are commonly used to represent motion of the backbone [17]. We note that these values depend on a number of other factors such as the overall resolution of the protein structure, crystal contacts, and applied refinement procedures [18, 19]. As a result, they are usually normalized as [20, 21]:

$$B_{i,normalized} = (B_i - B_{avg}) / \sigma$$

where B_{avg} is the average of the B-factor values of a given structure and σ is the corresponding standard deviation. The distribution of B-factor values along a protein sequence reflects flexibility and dynamics of the underlying structure. For instance, protein core is usually characterized by low B-factor values since it should be well packed to provide rigidity for the entire structure. At the same time, surface would usually include some flexible regions which would have high B-factor values. The reason is that the protein interacts with other molecules, which requires certain degree of structural flexibility.

To date, only a handful of methods performed sequence-based prediction of B-factor values. The first two methods used very simple techniques based on weighted sum [22] and averaging [17], which were followed by a method that used logistic regression predictor and a more complex feature representation based on the sequence composition and the predicted secondary structure [23]. The two recent methods encode the sequence with multiple alignment profile and use support vector regression (SVR) [24] and neural network [25] algorithms to predict the B-factors. The neural network method utilizes additional features based on the predicted secondary structure, solvent accessibility, and the secondary structure content. We note that several works reduce the real-value prediction to a classification problem in which the B-factor values are binned with a threshold into two categories that correspond to rigid and flexible residues, respectively [23, 25]. The main disadvantage of the latter methods is that the selection of the cutoff thresholds is neither objective nor optimal [24].

Solvent Accessibility and Residue Depth

The solvent-accessible surface area (ASA) of a residue indicates its level of burial (or exposure) to solvent in the protein structure and is usually expressed in terms of relative solvent accessibility (RSA). The RSA is defined as:

$$RSA_i = 100\% \times ASA_i / MASA_i$$

where ASA_i is the solvent-accessible surface area of the i^{th} residue observed in a given structure, and the $MASA_i$ is the maximum obtainable value of the solvent-accessible surface area for the corresponding type of the amino acid [26]. Alternatively, ASA value is normalized by the maximum value of exposed surface area obtained for an extended tripeptide conformation of Ala-X-Ala or Gly-X-Gly [27]. The reason for normalization is that different amino acids have different ASA distributions with largely varying mean and median values [28]. The RSA values range between 0 (fully buried) and 100% (fully exposed). The knowledge of solvent accessibility is useful in the context of understanding relationships between the structure and function of a protein because active sites are often located on the surface.

The sequence-based prediction of solvent accessibility was addressed by relatively large number of methods. The first method was developed in 2003 and applied a neural network algorithm for predictions [29]. Majority of the newer methods encode the protein sequence using multiple alignment and use a variety of prediction algorithms such as neural networks [30-33], look-up table [34], linear regression [35], multiple linear regression (MLR) [36], support vector machines (SVMs) [37], SVR [38], and two stage SVR [39, 40], and energy optimization [41]. Similarly, as in the case of the B-factor predictions, some of the existing RSA prediction methods cast the real-value prediction as a classification problem using a cutoff threshold that results in categorizing the residues as being either exposed or buried. These methods use a wide range of predictors including neural networks [26, 42-44], Bayesian statistics [45], substitution matrix [46, 47], information theory [48], multiple linear regression [49], SVM [50] and two-stage SVM [51]. However, the inherent problem with imposing the threshold is that it arbitrarily separates buried and exposed residues, and this arbitrariness is due to the extent of variations of RSA values for different residues [30].

Since the RSA values do not provide information how to characterize the buried residues, i.e., the ASA values of the buried residues are zero or near zero, an alternative descriptor was developed as follows. An atom depth is defined as the minimum distance between an atom and a dot of solvent-accessible surface [52], or the distance between the atom and its closest solvent-accessible neighbor [53], and finally based on volume [54]. The depth values can be used to identify a subset of residues that make the largest contribution to the stability of the molecule, i.e., residues at the protein core [52]. These residues are of special importance since burial of the core residues is a major driving force for folding [55]; recent research shows that the most deeply buried residues in the native protein fold might be the first to fold [56]. We note that so far only one method for prediction of residue depth, based on the SVR predictor that represents the input sequence using multiple sequence alignment, was developed [57].

Contact Number

The contact number (also called coordination number or Ooi number) of a given residue is defined as the number of C_α atoms of other (neighboring) residues within a sphere around the C_α atom of that given residue (that usually excludes the two nearest-neighbor residues along the sequence) [58, 59]. In a recent work, this definition was modified by considering C_β atoms (C_α for glycine) instead of C_α and smoothing the boundary of the sphere by a sigmoid function [60]:

$$O_i = \sum_{j:|j-i|>2} \sigma(r_{ij})$$

where r_{ij} is the distance between the C_β atoms of the i^{th} and j^{th} residues and:

$$\sigma(r) = \frac{1}{1 + e^{w(r-d_c)}}$$

where w is a constant determining the sharpness of the boundary of the sphere and d_c is a cutoff radius. The commonly accepted value of w is 3 [60-62]. The newer definition of the contact number results in a floating point number while the original definition produces an integer. The contact numbers provide a useful input to constrain de novo prediction of the tertiary protein structure. This is since the number of possible protein conformations that satisfy constraints imposed by the contact numbers along the protein sequence is very limited [63].

The contact number can be predicted from the protein sequence with the use of four methods. The first attempt was made in 1980, where a very simple empirical predictor was devised [58]. The other three methods use multiple alignment to represent the input sequence and linear regression [60], SVR [64], and then critical random networks [62] algorithms to do the predictions. The real value contact number was also reduced to two states that are defined by the values that are higher or lower than a threshold determined by the average value of the residue distribution. The corresponding classification was addressed by several prediction methods [44, 65, 66].

Residue-Wise Contact Order

The residue-wise contact order (RWCO) was developed based on several formerly developed structural descriptors. The relative contact order, which describes the complexity of protein topology, was used to study the correlation between protein topology and folding rate [67]. This descriptor was extended into residue contact order (RCO), which is the average contact order of a given residue [68]. The RWCO, a generalization of RCO, is a sum of sequence separations between the i^{th} residue and the contacting residues and is defined as [62, 69]:

$$RWCO_i = \sum_{j:|i-j|>2} |i - j| \, C_{ij}$$

where $C_{ij} = 1$ if the i^{th} and j^{th} residues are in contact and $C_{ij} = 0$ otherwise, and where the two nearest-neighbor residues along the sequence are excluded. The RWCO values are usually normalized by the length of the corresponding protein chain and they are usually smoothed with the use of the sigmoid function $\sigma(r)$. We observe that $RWCO_i = n \times RCO_i$ where n is the number of contacting residues with the i^{th} residue. The contact order is a global descriptor (concerns the entire protein), while RWCO and RCO are local (per-residue) descriptors. The usefulness of the RWCO was justified in a recent study in which it has been shown that the tertiary protein structure can be recovered from a set of three types of 1D descriptors which include the secondary structure, and the contact number and residue-wise contact order [61]. This in turn enabled design of novel methods for elucidating the sequence-structure relationship of proteins.

To date, three prediction methods were developed to predict RWCO values for protein residues using the sequence as the input. They all use multiple alignment while employing different prediction algorithms such as critical random networks [62], linear regression [69], and SVR [70].

BACKBONE TORSION ANGLES

The protein backbone consists of a linked sequence of rigid planar peptide groups. The rotational angle of the C–N bond is fixed at 180° for the common trans-conformation and 0° for the rare cis-conformation. As a result, the backbone can be described by two rotation angles (torsion angles) of the Cα–N bond (Φ) and the Cα–C bond (Ψ), which by convention vary between -180° and 180°. The distribution of the Φ/Ψ angles in protein structures is clustered around alpha (centered at $\Phi = -60°$ and $\Psi = -40°$), beta (centered at $\Phi = -120°$ and $\Psi = 120°$) and L-alpha (centered at $\Phi = 60°$ and $\Psi = 0°$) regions of the Ramachandran plot [71]. Analysis of the torsion angles shows that α-helices and β-sheets consist of residues with the angles distributed mostly in the alpha and beta Φ/Ψ angle regions, respectively [72]. This indicates that the knowledge of the torsion angles provides useful information for learning secondary protein structure.

Several methods have been developed for prediction of the Φ/Ψ angles. The first method, which predicts Ψ angles using a neural network in the context of improving the accuracy of secondary-structure prediction, was proposed in 2005 [73]. This was followed by another neural network method that predicts Ψ angles proposed in 2007 [33]. Recently, the neural network predictor was used to predict both Φ and Ψ angles [74]. As in the case of the B-factor, solvent accessibility, and contact number, several methods were developed to predict discrete dihedral-angle states [72, 75-81].

Secondary Structure Content

The secondary structure content is defined as the percentage of the α-helix, β-strand, and coil secondary structures in the protein sequence:

$$content_x = count_x / L$$

where $x = \{\alpha\text{-helix}, \beta\text{-strand}, \text{coil}\}$, $count_x$ denotes the number of residues assuming secondary structure of type x, and L is the length of the protein chain. Alternatively, instead of using the three secondary structure states, some methods address a finer division into secondary structures that include eight states defined with DSSP [82]. In this case $x = \{\alpha$-helix, β-strand, β-bridge, 3_{10}-helix, π-helix, H-bonded turn, bend, random coil$\}$. The content encapsulates the bulk (protein-wide) information concerning secondary structure without the knowledge of which residues assume a particular secondary structure. This information is useful to characterize an overall type of the protein fold, such as those defined in the SCOP [3] and CATH [4] databases.

The first attempt to predict the secondary structure content dates to 1970's when MLR method was used with amino acid composition of the protein sequence as the sequence representation [83]. Subsequent attempts used either neural networks or MLR methods and a variety of features computed from the protein sequence as input to the prediction method. The features include the molecular weight of a protein [84], auto-correlation functions based on hydrophobicity [85-87], pair-coupled composition [88-91], a selected subset

of composition vector features [92], composition moment vector [93], multiple alignment [94] and various physicochemical properties of amino acids combined with their composition [95]. One exception is a method that uses analytic vector decomposition predictor [96, 97]. Among the above-mentioned methods, five address eight-state content prediction [88-91, 94], while the remaining methods predict the three types of secondary structures. Several researchers investigated impact of a priori knowledge of structural classes on the quality of the content prediction [86, 98, 99]. The main drawback of the latter methods is that they require knowledge of the structural class of the input sequence. This information could be either inferred based on the known secondary structure, or predicted, but structural class prediction is difficult and is characterized by relatively low accuracy [100, 101].

The secondary structure content is closely related to structural classes, which categorize protein structures based on the amounts and arrangement of the constituent secondary structures. The three most commonly considered classes include all-α (proteins that contain mostly helices), all-β (proteins that contain mostly strands), and mixed class (proteins with both helices and strands), although several definition that consider different number of classes were proposed [3, 4, 97, 102, 103]. A recent survey summarizes and contrasts these definitions [104]. We note that the prediction of structural classes received a wide attention resulting in the development of numerous prediction methods [100, 102, 103, 105-122].

Folding Rate

The folding rate measures how fast a protein folds from the unfolded state to its native tertiary structure. Although the folding rates are sometimes measured in different experimental conditions, a recent contribution by Maxwell and colleagues established a set of standard conditions. They require 25°C, pH of 7.0, and 50nM buffer [123]. The folding rates are usually represented as decimal or natural logarithm of the protein folding rate in water, $\log(k_f)$, which are negatively correlated with the actual folding time.

A number of methods were developed to predict folding rates using protein sequence as the input. In the first attempt, a simple linear function of the effective chain length, which is computed using predicted secondary structure, was used to perform predictions [124]. Several other linear regression models that use features computed from physicochemical properties and composition of the constituent residues were recently developed [125-127]. Similarly, as in the case of the secondary structure content, a priori knowledge of structural classes was found to be useful in building the sequence-based predictors [128, 129]. A recent in-depth review of the methods used to predict folding rates can be found in [130].

COMPARISON OF SEQUENCE-BASED REAL-VALUE PREDICTION METHODS

In spite that the above-mentioned descriptors address a diverse range of structural aspects (such as the exposure to the solvent, spatial position and packing of the residues, their flexibility, and amount of secondary structures and folding time of a protein) the methods that address them share several similarities. We compare different designs that address real-value predictions based on the information how the input sequence is encoded and which prediction algorithms are used. We also investigate evaluation standards, which include benchmark datasets, test criteria, and procedures adopted for each of these tasks.

Table (**1**) presents a high-level overview of the real-value prediction methods. Over 50 real-value prediction methods were developed, with most of them published in the last five years, and in case of four descriptors the real values were collapsed into categorical predictions (which resulted in development of additional prediction methods discussed in the Background section). We observe that two descriptors, namely, solvent accessibility and secondary structure content, attracted the most attention. In some other cases, such as sequence-based prediction of backbone torsion angles and residue depth, the number of existing prediction methods is small and as the result they are excluded from further discussion. In the case of residue depth, the reason is that this task was defined very recently, while in the case of torsion angles the low count is due to an overlap with a wide variety of sequence-based secondary structure prediction methods. Our subsequent discussion also excludes folding rate prediction methods as they are discussed in depth in a recent review [130].

Table 1. Summary of Sequence-Based Real Value Prediction Methods

Scope of the Prediction	Prediction Target	# Published Methods	Year First Method was Published	Prediction of Discretized Target
Per Residue (local)	B-factor	5	1985	Yes
	Solvent accessibility	13	2003	Yes
	Residue depth	1	2008	No
	Contact number	4	1980	Yes
	Residue-wise contact order	3	2005	No
	Backbone torsion angles	3[1]	2005	Yes
Per Protein (global)	Secondary structure content	16[2]	1973	No
	Folding rate	6[3]	2004	No

[1] only one method predicts both Φ and Ψ angles; the remaining two methods predict only Ψ angle.
[2] 3 out of the 16 methods require the knowledge of structural classes.
[3] 2 out of the 6 methods require the knowledge of structural classes.

DESIGN OF THE PREDICTION METHODS

Table (**2**) compares the prediction methods with respect to the input sequence representation and the prediction algorithms used. For each descriptor (target), the corresponding prediction methods are ordered chronologically in the descending order.

From Table (**2**) we observe that:

- The information extracted directly from the sequence (such as composition vector, occurrence of sequence motifs, physicochemical and structural properties of amino acids) is often supplemented with additional information that includes multiple sequence alignment and results of other sequence-based prediction methods.
- One of the most popular inputs is multiple sequence alignment which is computed using PSI-BLAST algorithm [131]. The most commonly used output of the PSI-BLAST is the position-specific scoring matrix (PSSM), which is a 20 dimensional matrix (20 dimensions per each residue in the sequence) that provides log-odds scores for finding a particular matching amino acid in the target sequence.
- The inputs include results generated by other sequence-based prediction methods, such as predicted secondary structure and solvent accessibility. In particular, the predicted secondary structure was found useful in B-factor prediction, solvent accessibility prediction, and for prediction of residue-wise contact order. The prediction of the secondary structure was done with several methods that include PROFsec [132, 133], PHD [134], and PSI-PRED [135]. We observe that PSI-PRED was used in three out of five cases. The solvent accessibility, which was used in prediction of B-factor values, was predicted with the use of the PROFacc method [132, 133].
- The content prediction methods use the least amount of information. Only one prediction system uses multiple sequence alignment and none of the methods uses other predictions. This is because these methods address global (per sequence) predictions, while the PSI-BLAST and secondary structure are predicted per residue, which introduces a challenge with respect to the design of the corresponding input features. We observe that such features could be designed based on an approach reported in [100, 106].
- A relatively large variety of prediction algorithms was used. The most popular algorithms include support vector regression, neural network, and multiple linear regression. Two-stage predictors, which involve building a second prediction model that takes as an input a set of predictions provided by prediction model incorporated in the first stage, were developed only for solvent accessibility. The second stage model uses several neighboring predictions produced by the first stage to improve the results. We believe that such design can prove useful when using other per-residue prediction targets.
- A side-by-side comparison of the content prediction methods shows that support vector regression generates better results than neural networks and multiple linear regression for the secondary structure content prediction [94]. Similarly, support vector regression is shown comparable or better than linear regression and critical random networks for the RWCO prediction [70]. In case of other descriptors a direct comparison of different algorithms is more difficult and subjective due to the use of different datasets, evaluation procedures, and overall lack of comprehensive comparisons; see the next section for details.

EVALUATION PROTOCOLS

Table (**3**) summarizes evaluation protocols used in developing real-value prediction methods.

The prediction methods were evaluated with two main types of tests, out-of-sample and in-sample. The in-sample tests (i.e., resubstitution) are based on the protein sequences used to develop the prediction model, while out-of-sample tests (i.e., cross validation and single-split) are based on using chains that were not used to design the model. The cross validation tests divide the dataset into n subsets and use n-1 subsets to generate the model and the remaining subset to evaluate it; this is repeated n times, each time using a different subset as the test set. The two most popular cross validation tests include the case when n is a small constant (usually 3, 5, or 10), and when n is equal to the number of chains in the dataset (jackknife test).

From Table (**3**) we note that:

- Some predictors are characterized by the lack of standard benchmark datasets, i.e., each new predictor is evaluated on a different dataset. This is true in the case of B-factor and contact number prediction methods. The same is true for the prediction of solvent accessibility and secondary structure content. Although in this case some benchmark datasets exist, their number is relatively large. More specifically, the total of ten datasets (nine are benchmark datasets) and twenty datasets (four benchmark datasets) were used to evaluate solvent accessibility and content predictors, respectively. This makes it difficult to establish a relative quality of prediction systems and constitutes a substantial challenge for developers of new methods (as they should be compared with existing methods). The only exception is the prediction of residue-wise contact order where the three developed methods are tested on the same dataset.
- The datasets vary in size. The smallest datasets include several sequences, while the largest include several thousands chains.
- Most datasets were established using filtering based on the maximal pairwise sequence similarity, while authors used several different identity thresholds. They vary between 22% and 50%, with 25% being the most frequently used value. The choice of the thresholds is motivated by the fact that sequences with low similarity are more difficult to predict. For instance, more than 95% of protein chains characterized by the 20-25% pairwise identity (referred to as the twilight-zone similarity) have different struc

Table 2. Comparison of the Input Sequence Representations and Prediction Algorithms Used to Address Prediction of B-factor, Solvent Accessibility, Contact Number, Residue-Wise Contact Order, and Secondary Structure Content.

Prediction Target	Input (Sequence Representation)	Prediction Algorithm[1]	Reference
B-factor	sequence multiple sequence alignment	SVR	[24]
	sequence multiple sequence alignment predicted 2-state solvent accessibility and fraction of surface residues predicted secondary structure and secondary structure content	NN	[25]
	sequence K2 entropy predicted secondary structure	LR	[23]
	sequence	Sliding window averaging	[17]
	sequence	Weighted sum	[22]
Solvent Accessibility	multiple sequence alignment sequence predicted secondary structure	Two-stage SVR	[40]
	multiple sequence alignment entropy	SVM	[37]
	multiple sequence alignment sequence physicochemical and structural properties of amino acids	NN	[33]
	multiple sequence alignment	Two-stage SVR	[39]
	multiple sequence alignment	NN	[32]
	sequence	n/a (energy optimization)	[41]
	multiple sequence alignment	MLR	[36]
	multiple sequence alignment predicted secondary structure	NN	[31]
	multiple sequence alignment	SVR and LinR	[35]
	sequence	SVR	[38]
	sequence	Look-up table	[34]
	multiple sequence alignment	NN	[30]
	sequence	NN	[29]
Contact Number	multiple sequence alignment	Critical random network	[62]
	sequence multiple sequence alignment	LinR	[60]
	sequence multiple sequence alignment	SVR	[64]
	sequence	n/a	[58]
Residue-Wise Contact Order	sequence multiple sequence alignment predicted secondary structure	SVR	[70]
	sequence multiple sequence alignment	LinR	[69]
	multiple sequence alignment	Critical random networks	[62]

Table 2. Contd….

Prediction Target	Input (Sequence Representation)	Prediction Algorithm[1]	Reference
Secondary Structure Content	sequence physicochemical and structural properties of amino acids	MLR	[95]
	sequence	SVR	[91]
	multiple sequence alignment	SVR	[94]
	sequence	NN	[93]
	sequence	MLR	[92]
	sequence	NN	[90]
	sequence and hydrophobicity	MLR	[87]
	sequence hydrophobicity	MLR	[86]
	sequence	MLR	[89]
	sequence	MLR	[88]
	sequence and hydrophobicity	MLR	[85]
	sequence	Vector decomposition	[96, 97]
	sequence	NN	[84]
	sequence	MLR	[83]

[1]SVR (support vector regression); NN (neural network); LR (logistic regression); LinR (linear regression); MLR (multiple linear regression); SVM (support vector machine); n/a means that the prediction was performed without the use of a prediction algorithm (using an empirical model).

tures [136], which poses a substantial challenge for higher-level (secondary and tertiary) structure prediction methods. For instance, the accuracy of the secondary structure prediction methods trained and tested on a protein set in which any pair of sequences shares twilight-zone similarity drops to only 65-68% [137]; when higher similarity is present the accuracy rises to above 80% [138]. Only in the case of the RWCO and contact number predictions the filtering is done based on homology (one chain per superfamily).

- The most commonly used test procedures are based on cross validation, although we observe that no standards are imposed which specific one to use. This again makes it very difficult to perform comparison between different methods. The number of cross validation folds ranges from 3 to 10, while in content prediction some authors use the jackknife (leave-one-out) test. The jackknife test is computationally expensive, which prevents its use for performing evaluation of per-residue (local) predictors.
- The prediction quality was measured by a number of criteria that include Pearson correlation coefficient (PCC), mean absolute error (MAE), normalized mean absolute error, absolute deviation, average relative deviation, root mean square error, and standard error. All of these criteria are computed between the predicted and the actual (true) values of the corresponding descriptors. The most commonly used are PCC and MAE.

APPLICATIONS

The motivation to develop the above-mentioned prediction systems stems from the benefits provided by the knowledge of the corresponding real-value descriptors. In this section we briefly summarize applications in which the above-discussed descriptors were used.

B-factor

The knowledge of B-factor values was used in prediction of protein flexibility [17, 22], analysis of protein thermal stability [139, 140] and active sites [141-143], correlating the side chain mobility with protein conformation [20, 144], analysis of disordered regions [23, 145] and protein dynamics [146], prediction of protein-protein binding sites [147], and analysis of evolutionary divergence of protein backbone dynamics [148] and enzymatic reactions [149].

Solvent Accessibility and Residue Depth

The solvent accessibility was used for tertiary structure prediction and fold recognition [150, 151], to develop amino acid substitution matrix [152], to predict stability of protein mutants [153, 154] to predict protein-protein interaction sites [155, 156], protein interfaces [157], protein domains [158], transmembrane domains [159], long-range contacts [160], and residue contacts [161]. Solvent accessibility and an accurate estimation of the solvent accessible surface were also found important for studying protein-protein binding interaction and the low-frequency collective motion in biomacromolecules [162, 163]. When compared with the solvent accessibility, the residue depth is characterized by higher correlation with residue conservation, which motivated a number of interesting applications. The residue depth was used to analyze amide hydrogen/deuterium exchange rates in nuclear magnetic resonance experiments [164], local packing arrangements in the protein core [165], to analyze and predict functional sites such as catalytic sites of enzyme [166] and phosphorylation sites [53, 167], and to perform prediction of the protein folds [151, 168, 169].

Table 3. Comparison of the Datasets, Test Procedures, and Test Criteria Used to Evaluate Methods for Prediction of B-factor, Solvent Accessibility, Contact Number, Residue-Wise Contact Order, and Secondary Structure Content

Prediction Target	Datasets[1]	Test Procedure[2]	Test Criteria[3]	Reference
B-factor	766 chains (identity<25%)	5-fold CV	PCC	[24]
	1513 chains (identity<22%)	3-fold CV	PCC	[25]
	290 chains (identity<25%)	30 random single-split tests	PCC	[23]
	92 chains	resubstitution	PCC	[17]
	31 chains (identity<50%)	resubstitution	PCC	[22]
Solvent Accessibility	215 chains *Manesh215* (identity<25%)	single-split 5-fold CV	MAE PCC	[40]
	126 chains *RS126* 480 chains *Barton480* (identity<25%)	5-fold CV	MAE PCC	[37]
	2640 chains (identity<25%)	10-fold CV	MAE PCC	[33]
	215 chains *Manesh215* (identity<25%) 338 chains *Carugo338* (identity<25%) 480 chains *Barton480* (identity<25%)	3-fold CV	MAE PCC	[39]
	480 chains *Barton480* (identity<25%)	single-split	MAE	[32]
	126 chains *RS126* 215 chains *Manesh215* (identity<25%) 338 chains *Carugo338* (identity<25%) 480 chains *Barton480* (identity<25%)	independent test	PCC	[41]
	480 chains *Barton480* (identity<25%) 1277 chains *Yuan1277* (identity<25%)	5-fold CV	MAE PCC	[36]
	215 chains *Manesh215* (identity<25%) 480 chains *Barton480* (identity<25%)	5-fold CV	MAE PCC	[31]
	135 chains *S135* (identity<50%, e-value<0.001) 149 chains *S149* (identity<50%, e-value<0.001) 156 chains *S156* (identity<50%, e-value<0.001) 163 chains *S163* (identity<50%, e-value<0.001)	10-fold CV	MAE PCC	[35]
	480 chains *Barton480* (identity<25%)	3-fold CV	MAE PCC	[38]
	480 chains *Barton480* (identity<25%)	JK	MAE PCC	[34]
	135 chains *S135* (identity<50%, e-value<0.001) 149 chains *S149* (identity<50%, e-value<0.001) 156 chains *S156* (identity<50%, e-value<0.001) 163 chains *S163* (identity<50%, e-value<0.001)	3-fold CV	MAE PCC	[30]
	126 chains *RS126* 215 chains *Manesh215* (identity<25%) 338 chains *Carugo338* (identity<25%) 480 chains *Barton480* (identity<25%)	3-fold CV	MAE PCC	[29]
Contact Number	680 chains *KN680* (one sequence per superfamily)	15 random single-split tests	PCC DevA	[62]
	1050 chains (identity<30%)	10 random single-split tests	PCC DevA DevR	[60]
	945 chains (identity<25%)	3-fold CV (2 folds to test & 1 to train)	PCC RMSE	[64]
	39 chains	single-split	PCC	[58]

Table 3. Contd….

Prediction Target	Datasets[1]	Test Procedure[2]	Test Criteria[3]	Reference
Residue-wise Contact Order	680 chains *KN680* (one sequence per superfamily)	15 random single-split tests	PCC DevA RMSE	[70]
	680 chains *KN680* (one sequence per superfamily)	15 random single-split tests	PCC DevA	[69]
	680 chains *KN680* (one sequence per superfamily)	15 random single-split tests	PCC DevA	[62]
Secondary Structure Content	2187 chains (identity<25%) 2483 chains (identity<40%)	10-fold CV resubstitution	MAE NMAE DevA	[95]
	202 chains *C202* (identity<35%) 244 chains *C244* (identity<35%) 513 chains (SD score≥5)	7-fold CV independent test resubstitution	MAE	[91]
	202 chains *C202* (identity<25%) 5796 chains (identity<40%)	single-split	MAE	[94]
	11206 chains 2439 chains (identity<25%)	single-split JK	MAE DevA	[93]
	475 chains *E475* (identity≤35%)	single-split	MAE SE	[92]
	202 chains *C202* (identity<25%) 244 chains *C244* (identity<25%)	independent test resubstitution	MAE DevA	[90]
	740 chains (identity<30%)	resubstitution JK	MAE DevA	[87]
	210 chains 143 chains	resubstitution JK	MAE DevA	[86]
	628 chains (identity<25%) 52 chains (identity<35%)	independent test resubstitution	MAE DevA	[89]
	202 chains *C202* (identity<35%) 244 chains *C244* (identity<35%)	independent test resubstitution	MAE DevA	[88]
	262 chains *E262* (identity<35%) 347 chains	independent test resubstitution JK	MAE DevA	[85]
	166 chains (identity<35%) 262 chains *E262* (identity<35%) 398 chains (identity<35%) 475 chains *E475* (identity≤35%)	resubstitution JK	MAE DevA PCC	[96, 97]
	104 chains 15 chains (identity≤33%)	resubstitution independent test	MAE DevA	[84]
	18 chains	resubstitution JK	PCC MAE	[83]

[1] We list the number of used protein chains (italics denote name of a benchmark dataset, i.e., dataset used across multiple contributions); text in brackets specifies whether the chains were filtered based on sequence similarity or homology.

[2] CV (cross validation); JK (jackknife); single-split means that the original dataset was split into training and tests sets; resubstitution means that the model was tested on the dataset used to design the prediction method; independent test means that the predictor was designed using another dataset and tested on the listed dataset(s).

[3] PCC (Pearson correlation coefficient); MAE (mean absolute error); NMAE (normalized mean absolute error); DevA (absolute deviation); DevR (average relative deviation); RMSE (root mean square error); SE (standard error).

Contact Number and Residue-Wise Contact Order

Knowledge of the number of residue contacts is important for deriving constraints to be used in modeling protein folding, protein structure, and in scoring remote homology searches [170-173] and in assessing the quality of protein structures in protein structure prediction [174]. Given the contact numbers along the protein chain, the number of possible protein conformations that satisfy these constraints was shown to be limited [63]. These constrains were used in mo-

lecular dynamics simulations [61, 175] and simulations of lattice proteins [63]. A similar restraint was used for off-lattice simulations [176]. The usefulness of the contact number and residue-wise contact order were justified in a recent study which shows that the tertiary protein structure can be recovered from a set of three one-dimensional descriptors that include secondary structure, contact number, and residue-wise contact order [61].

Backbone Torsion Angles

The knowledge of torsion angles have been used to improve fold recognition [72], sequence alignment [177], and accuracy of secondary structure prediction [73, 80].

Secondary Structure Content

Predicted secondary structure content was used in several areas such as structural class prediction [110, 178] and analysis of interactions between CapZ protein and cell membranes [179]. In some applications, the predicted secondary structure and true (actual) secondary structure were used to compute the content. These content values were used in analysis of prion proteins [180], prediction of coding and noncoding RNAs [181], folding rates [124, 130, 182], folding transition-state position [183], and enzyme class [184], and to distinguish between enzyme and non-enzyme proteins [185].

CONCLUSIONS

The real-value prediction methods that address structural aspects of proteins span a wide spectrum of descriptors such as solvent accessibility, residue depth, B-factors, contact numbers, residue-wide contact numbers, backbone torsion angles, secondary structure content, and folding rates. In this review we summarized and compared prediction methods from two important perspectives, their design and their evaluation procedures. We show that in spite of addressing different objectives a number of similarities exist that can be exploited in development of new prediction methods.

The most popular prediction algorithms include support vector regression, neural networks, and multiple linear regression. The two-stage design was found useful in solvent accessibility prediction, while no attempts were made to use such design in other tasks that address predictions for individual amino acids. The input to the prediction algorithms is computed either directly from the protein sequence, and/or from other sequence-derived sources such as multiple sequence alignment or results of other sequence-based prediction methods, including predicted secondary structure and solvent accessibility. One of the very popular inputs is the PSSM matrix generated by the PSI-BLAST algorithm. In spite of the similarities in the design, the only method that predicts several descriptors at the same time is Real-SPINE [33]. It predicts solvent accessibility and backbone Ψ dihedral angles. We believe that similarities between the real-value predictors should be exploited to develop more methods such as Real-SPINE. For instance, contact number and solvent accessibility are characterized by negative correlation with each other [60] that could be used in new designs.

The evaluation procedures use a variety of datasets, test procedures, and quality indices. We observe the overall lack of well-established benchmark datasets which makes it difficult to compare the relative quality of different prediction systems; it also poses a big challenge to developers of new predictors. Most of the used datasets were created based on the pairwise sequence similarity filters. The maximal identity thresholds vary between 22% and 50%, with the value of 25% being the most frequently used. The evaluations are mostly based on cross validation tests with the number of folds ranging from 3 to 10. Although the prediction quality is measured using seven different indices, the two most popular ones are Pearson correlation coefficient and mean absolute error.

The breath and number of the discussed applications for the predicted real-values, which concern analysis and predictions of various structural and functional aspects of proteins, signify the importance of the real-value prediction methods. We believe that these applications call for further development of more accurate and integrated prediction methods.

REFERENCES

[1] Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* **2000**; 28(1): 235-42.

[2] Boeckmann B, Bairoch A, Apweiler R, *et al*. The swiss-prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **2003**; 31: 365-70.

[3] Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acid Res* **2004**; 32: D226-9.

[4] Pearl FMG, Lee D, Bray JE, *et al*. Assigning genomic sequences to CATH. *Nucleic Acids Res* **2000**; 28(1): 277-82.

[5] Pruitt KD, Tatusova T, Maglott DR. (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **2005**; 33: D501-4.

[6] Chen K, Jiang Y, Du L, Kurgan L. Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J Comp Chem* **2008**; DOI: 10.1002/jcc.21053

[7] Chou KC, Shen HB. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Comm* **2007**; 360: 339-45.

[8] Pu X, Guo J, Leung H, Lin Y. Prediction of membrane protein types from sequences and position-specific scoring matrices. *J Theor Biol* **2007**; 247(2): 259-65.

[9] Shen HB, Yang J, Chou KC. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *J Theor Biol* **2006**; 240: 9-13.

[10] Shen HB, Chou KC. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochem Biophys Res Comm* **2005**; 334: 288-92.

[11] Chou KC, Cai YD. Using GO-PseAA predictor to identify membrane proteins and their types. *Biochem Biophys Res Comm* **2005**; 327: 845-7.

[12] Chou KC, Cai YD. Prediction of membrane protein types by incorporating amphipathic effects. *J Chem Inf Mod* **2005**; 45: 407-13.

[13] Shen HB, Chou JJ. MemBrain: Improving the accuracy of predicting transmembrane helices. *PLoS One* **2008**; 3(6): e2399.

[14] Ganapathiraju M, Balakrishnan N, Reddy R, Klein-Seetharaman J. Transmembrane helix prediction using amino acid property features and latent semantic analysis. *BMC Bioinformatics* **2008**; 9 Suppl 1: S4.

[15] Martin-Galiano AJ, Smialowski P, Frishman D. Predicting experimental properties of integral membrane proteins by a naive Bayes approach. *Proteins* **2008**; 70(4): 1243-56.

[16] Blow D. Outline of crystallography for biologists, New York: Oxford University Press, 2002; 237.

[17] Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. *Proteins* **1994**; 19: 141-9.

[18] Sheriff S, Hendrickson WA, Stenkamp RE, Sieker LC, Jensen LH. Influence of solvent accessibility and intermolecular contacts on atomic mobilities in hemerythrins. *Proc Natl Acad Sci USA* **1985**; 82: 1104-7.

[19] Tronrud DE. Knowledge-based B-factor restraints for the refinement of proteins. *J Appl Crystallogr* **1996**; 29: 100-4.
[20] Carugo O, Argos P. Correlation between side chain mobility and conformation in protein structures. *Protein Eng* **1997**; 10: 777-87.
[21] Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. *Protein Sci* **2003**; 12: 1060-72.
[22] Karplus PA, Schulz GE. Prediction of chain flexibility in proteins - a tool for the selection of peptide antigens. *Naturwissenschaften* **1985**; 72: 212-3.
[23] Radivojac P, Obradovic Z, Smith DK, *et al*. Protein flexibility and intrinsic disorder. *Protein Sci* **2004**; 13: 71-80.
[24] Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. *Proteins* **2005**; 58: 905-12.
[25] Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. *Proteins* **2005**; 61: 115-26.
[26] Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* **1994**; 20: 216-26.
[27] Samanta U, Bahadur RP, Chakrabarti P. Quantifying the accessible surface area of protein residues in their local environment. *Protein Eng* **2002**; 15: 659-67.
[28] Lins L, Thomas A, Brasseur R. Analysis of accessible surface of residues in proteins. *Protein Sci* **2003**; 12: 1406-17.
[29] Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins* **2003**; 50(4): 629-35.
[30] Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* **2004**; 56(4): 753-67.
[31] Garg A, Kaur H, Raghava GP. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins* **2005**; 61(2): 318-24.
[32] Araúzo-Bravo MJ, Ahmad S, Sarai A. Dimensionality of amino acid space and solvent accessibility prediction with neural networks. *Comput Biol Chem* **2006**; 30(2): 160-8.
[33] Dor O, Zhou Y. Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. *Proteins* **2007**; 68(1): 76-81.
[34] Wang JY, Ahmad S, Gromiha MM, Sarai A. Look-up tables for protein solvent accessibility prediction and nearest neighbor effect analysis. *Biopolymers* **2004**; 75(3): 209-16.
[35] Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol* **2005**; 12(3): 355-69.
[36] Wang JY, Lee HM, Ahmad S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins* **2005**; 61(3): 481-91.
[37] Wang JY, Lee HM, Ahmad S. SVM-Cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine. *Proteins* **2007**; 68(1): 82-91.
[38] Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. *Proteins* **2004**; 57(3): 558-64.
[39] Nguyen MN, Rajapakse JC. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins* **2006**; 63(3): 542-50.
[40] Chen K, Kurgan M, Kurgan L. Sequence Based Prediction of Relative Solvent Accessibility Using Two-stage Support Vector Regression with Confidence Values. *J Biom. Sci Eng* **2008**; 1: 1-9.
[41] Xu Z, Zhang C, Liu S, Zhou Y. QBES: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization. *Proteins* **2006**; 63(4): 961-6.
[42] Cuff JA, Barton GJ. Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **1999**; 40: 502-11.
[43] Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* **2002**; 18: 819-24.
[44] Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* **2002**; 47: 142-53.
[45] Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins* **1996**; 25: 38-47.
[46] Richardson CJ, Barlow DJ. The bottom line for prediction of residue solvent accessibility. *Protein Eng* **1999**; 12: 1051-4.
[47] Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. *Protein Eng* **2000**; 13(9): 607-9.
[48] Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA. Prediction of protein surface accessibility with information theory. *Proteins* **2001**; 42: 452-9.
[49] Li X, Pan X-M. New method for accurate prediction of solvent accessibility from protein sequence. *Proteins* **2001**; 42(1): 1-5.
[50] Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins* **2004**; 54: 557-62.
[51] Nguyen MN, Rajapakse JC. Prediction of protein relative solvent accessibility with a two-stageSVMapproach. *Proteins* **2005**; 59: 30-7.
[52] Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* **1999**; 7: 723-32.
[53] Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. *Biophys J* **2003**; 84: 2553-61.
[54] Varrazzo D, Bernini A, Spiga O, *et al*. Three-dimensional computation of atom depth in complex molecular structures. *Bioinformatics* **2005**; 21(12): 2856-60.
[55] Chan HS, Dill KA. Origins of structures in globular proteins. *Proc Natl Acad Sci USA* **1990**; 87: 6388-92.
[56] Pintar A, Pongor S. The "first in-last out" hypothesis on protein folding revisited. *Proteins* **2005**; 60: 584-90.
[57] Yuan Z, Wang Z-X. Quantifying the relationship of protein burying depth and sequence. *Proteins* **2008**; 70: 509-16.
[58] Nishikawa K, Ooi T. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int J Pept Protein Res* **1980**; 16: 19-32.
[59] Nishikawa K, Ooi T. Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J Biochem* **1986**; 100: 1043-7.
[60] Kinjo AR, Horimoto K, Nishikawa K. Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins* **2005**; 58: 158-65.
[61] Kinjo AR, Nishikawa K. Recoverable one-dimensional encoding of protein three-dimensional structures. *Bioinformatics* **2005a**; 21: 2167-70.
[62] Kinjo AR, Nishikawa K. Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structures from amino acid sequences using critical random networks. *Biophysics* **2005b**; 1: 67-74.
[63] Kabakcioglu A, Kanter I, Vendruscolo M, Domany E. Statistical properties of contact vectors. *Phys Rev E* **2002**; 65: 041904.
[64] Yuan Z. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics* **2005**; 6: 248.
[65] Pollastri G, Baldi P, Fariselli P, Casadio R. Improved prediction of the number of residue contacts in proteins by recurrent neural networks. *Bioinformatics* **2001**; (Suppl 1): S234-42.
[66] Fariselli P, Casadio R. Prediction of the number of residue contacts in proteins. *Proc Int Cort Intell Syst Mol Bio* **2000**; 146-151.
[67] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* **1998**; 277: 985-94.
[68] Kihara D. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci* **2005**; 14: 1955-63.
[69] Kinjo AR, Nishikawa K. Predicting residue-wise contact orders of native protein structure from amino acid sequence **2005c**; arXiv.org. q-bio.BM/0501015.
[70] Song JN, Burrage K. Predicting residue-wise contact orders in proteins by support vector regression. *BMC Bioinformatics* **2006**; 7: 425.
[71] Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **1963**; 7: 95-9.
[72] Kuang R, Leslie CS, Yang AS. Protein backbone angle prediction with machine learning approaches. *Bioinformatics* **2004**; 20(10): 1612-21.
[73] Wood MJ, Hirst JD. Protein secondary structure prediction with dihedral angles. *Proteins* **2005**; 59: 476-81.
[74] Xue B, Dor O, Faraggi E, Zhou Y. Real-value prediction of backbone torsion angles. *Proteins* **2008**; 72(1): 427-33.
[75] Kang HS, Kurochkina NA, Lee B. Estimation and use of protein backbone angle probabilities. *J Mol Biol* **1993**; 229: 448-60.

[76] Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* **2000**; 301: 173-90.

[77] deBrevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **2000**; 41: 271-87.

[78] Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* **2003**; 51: 504-14.

[79] deBrevern AG, Benros C, Gautier R, Valadie H, Hazout S, Etchebest C. Local backbone structure prediction of proteins. *In Silico Biol* **2004**; 4: 31.

[80] Mooney C, Vullo A, Pollastri G. Protein structural motif prediction in multidimensional - space leads to improved secondary structure prediction. *J Comput Biol* **2006**; 13: 1489-502.

[81] Zimmermann O, Hansmann UHE. Support vector machines for prediction of dihedral angle regions. *Bioinformatics* **2006**; 22: 3009-15.

[82] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**; 22(12): 2577-637.

[83] Krigbaum WR, Knutton SP. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. *Proc Natl Acad Sci USA* **1973**; 70: 2809-13.

[84] Muskal SM, Kim SH. Predicting protein secondary structure content. A tandem neural network approach. *J Mol Biol* **1992**; 225: 713-27.

[85] Zhang CT, Lin ZS, Zhang Z, Yan M. Prediction of helix/strand content of globular proteins based on their primary sequences. *Protein Eng* **1998a**; 11: 971-9.

[86] Zhang Z, Sun Z, Zhang CT. A new approach to predict the helix/strand content of globular proteins. *J Theor Biol* **2001**; 208: 65-78.

[87] Lin Z, Pan X. Accurate prediction of protein secondary structural content. *J Protein Chem* **2001**; 20: 217-20.

[88] Chou KC. Using pair-coupled amino-acid composition to predict protein secondary structure content. *J Protein Chem* **1999**; 18: 473-80.

[89] Liu W, Chou KC. Protein secondary structural content prediction. *Protein Eng* **1999**; 12: 1041-50.

[90] Cai YD, Liu XJ, Chou KC. Prediction of protein secondary structure content by artificial neural network. *J Comp Chem* **2003**; 24: 727-31.

[91] Chen C, Tian Y, Zou X, Cai P, Mo J. Prediction of protein secondary structure content using support vector machine. *Talanta* **2007**; 71(5): 2069-73.

[92] Pilizota T, Lucic B, Trinajstic N. Use of variable selection in modeling the secondary structural content of proteins from their composition of amino acid residues. *J Chem Inf Comp Sciences* **2004**; 44(1): 113-21.

[93] Ruan J, Wang K, Yang J, Kurgan L, Cios KJ. Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. *Artif Intel Med* **2005**; 35: 19-35.

[94] Lee S, Lee B, Kim D. Prediction of protein secondary structure content using amino acid composition and evolutionary information. *Proteins* **2006**; 62: 1107-14.

[95] Homaeian L, Kurgan L, Ruan J, Cios KJ, Chen K. Prediction of protein secondary structure content for the twilight zone sequences. *Proteins* **2007**; 69: 486-98.

[96] Eisenhaber F, Imperiale F, Argos P, Frommel C. Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. *Proteins* **1996a**; 25(2): 157-68.

[97] Eisenhaber F, Frommel C, Argos P. Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins* **1996b**; 25(2): 169-79.

[98] Zhang CT, Zhang Z, He Z. Prediction of the secondary structure of globular proteins based on structural classes. *J Prot Chem* **1996**; 15: 775-86.

[99] Zhang CT, Zhang Z, He Z. Prediction of the secondary structure contents of globular proteins based on three structural classes. *J Prot Chem* **1998b**; 17: 261-72.

[100] Kurgan L, Cios KJ, Chen K. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinformatics* **2008b**; 9: 226.

[101] Chen K, Kurgan L. PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics* **2007**; 23(21): 2843-50.

[102] Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. *J Biochem* **1986**; 99: 152-62.

[103] Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins* **1995**; 21: 319-44.

[104] Kurgan L, Zhang T, Zhang H, Shen S, Ruan J. Secondary structure based assignment of the protein structural classes. *Amino Acids* **2008a**; DOI: 10.1007/s00726-008-0080-3

[105] Zhang TL, Ding YS, Chou KC. Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. *J Theor Biol* **2008**; 250: 186-93.

[106] Chen K, Kurgan L, Ruan J. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J Comput Chem* **2008a**; 29: 1596-604.

[107] Chen C, Chen LX, Zou XY, Cai PX. Predicting protein structural class based on multi-features fusion. *J Theor Biol* **2008b**; 253: 388-92.

[108] Xiao X, Lin WZ, Chou KC. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. *J Comp Chem* **2008**; 29: 2018-24.

[109] Ding YS, Zhang TL, Chou KC. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. *Protein Pept Let* **2007**; 14: 811-15.

[110] Kurgan L, Chen K. Prediction of protein structural class for the twilight zone sequences. *Biochem Biophys Res Commun* **2007**; 357: 453-60.

[111] Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB. Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. *Biophys Chem* **2007**; 128: 87-93.

[112] Lin H, Li Q.Z. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. *J Comp Chem* **2007**; 28: 1463-6.

[113] Xiao X, Shao SH, Huang ZD, Chou KC. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. *J Comp Chem* **2006**; 27: 478-82.

[114] Niu B, Cai YD, Lu WC, Zheng GY, Chou KC. Predicting protein structural class with AdaBoost learner. *Protein Pept Let* **2006**; 13: 489-92.

[115] Kurgan, L, Homaeian, L. Prediction of structural classes for protein sequences and domains - impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. *Pattern Recognit* **2006**; 39: 2323-43.

[116] Kedarisetti KD, Kurgan L, Dick S. Classifier ensembles for protein structural class prediction with varying homology. *Biochem Biophys Res Commun* **2006**; 348: 981-8.

[117] Chen C, Tian YX, Zou XY, Cai PX, Mo JY. Using pseudo-amino acid composition and support vector machine to predict protein structural class. *J Theor Biol* **2006a**; 243: 444-8.

[118] Chen C, Zhou X, Tian Y, Zou X, Cai P. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal Biochem* **2006b**; 357: 116-21.

[119] Chou KC, Maggiora GM. Domain structural class prediction. *Protein Eng* **1998**; 11: 523-38.

[120] Chou KC, Liu W, Maggiora GM, Zhang CT. Prediction and classification of domain structural classes. *Proteins* **1998**; 31: 97-103.

[121] Chou KC, Zhang CT. Prediction of protein structural classes. *Crit Rev Biochem Mol Biol* **1995**; 30: 275-349.

[122] Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J Biol Chem* **1994**; 269: 22014-20.

[123] Maxwell KL, Wildes D, Zarrine-Afsar A. Protein folding: Defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci* **2005**; 14: 602-16.

[124] Ivankov DN, Finkelstein AV. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc Natl Acad Sci USA* **2004**; 101: 8942-4.

[125] Huang JT, Tian, J. Amino acid sequence predicts folding rate for middle-size two-state proteins. *Proteins* **2006**; 63(3): 551-4.

[126] Ma BG, Guo JX, Zhang HY. Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. *Proteins* **2006**; 65: 362-72.
[127] Huang LT, Gromiha MM. Analysis and prediction of protein folding rates using quadratic response surface models. *J Comput Chem* **2008**; 29(10): 1675-83.
[128] Gromiha MM. A statistical model for predicting protein folding rates from amino acid sequence with structural class information. *J Chem Inf Model* **2005**; 45: 494-501.
[129] Gromiha MM, Thangakami AM, Selveraj S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acid Res* **2006**; 34: W70-4.
[130] Gromiha MM, Selvaraj S. Bioinformatics approaches for understanding and predicting protein folding rates, *Curr Bioinformatics* **2008**; 3: 1-9.
[131] Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**; 25(17): 3389-402.
[132] Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth Enzymol* **1996**; 266: 525-39.
[133] Rost B. Protein secondary structure prediction continues to rise. *J Struct Biol* **2001**; 134: 204-18.
[134] Rost B, Sander C, Schneider R. PHD - an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci* **1994**; 10(1): 53-60.
[135] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **1999**; 292(2): 195-202.
[136] Rost B. Twilight zone of protein sequence alignments. *Protein Eng* **1999**; 2: 85-94.
[137] Lin K, Simossis V, Taylor W, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* **2005**; 21: 152-9.
[138] Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* **2006**; 7: 301.
[139] Vihinen M. Relationship of protein flexibility to thermostability. *Protein Eng* **1987**; 1: 477-80.
[140] Parthasarathy S, Murthy MRN. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng* **2000**; 13: 9-13.
[141] Carugo O, Argos P. Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins* **1998**; 31: 201-13.
[142] Yuan Z, Zhao J, Wang ZX. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng* **2003**; 16: 109-14.
[143] Mohan S, Sinha N, Smith-Gill J. Modeling the binding sites of anti-hen egg white lysozyme antibodies HyHEL-8 and HyHEL-26: an insight into the molecular basis of antibody cross-reactivity and specificity. *Biophys J* **2003**; 85: 3221-36.
[144] Eyal E, Najmanovich R, Edelman M, Sobolev V. Protein side-chain rearrangement in regions of point mutations. *Proteins* **2003**; 50: 272-82.
[145] Altman R, Hughes C, Zhao D, Jardetsky O. Compositional characteristics of relatively disordered regions in proteins. *Protein Pept Lett* **1994**; 1: 120-7.
[146] Navizet I, Lavery R, Jernigan RL. Myosin flexibility: Structure domains and collective vibrations. *Proteins* **2004**; 54: 384-93.
[147] Chung JL, Wang W, Bourne PE. Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins* **2006**; 62(3): 630-40.
[148] Maguid S, Fernández-Alberti S, Parisi G, Echave J. Evolutionary conservation of protein backbone flexibility. *J Mol Evol* **2006**; 63(4): 448-57.
[149] Fontana A, Spolaore B, Mero A, Veronese FM. Site-specific modification and PEGylation of pharmaceutical proteins mediated by transglutaminase. *Adv Drug Deliv Rev* **2008**; 60(1): 13-28.
[150] Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol* **1997**; 267(4): 1026-38.
[151] Liu S, Zhang C, Liang S, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. *Proteins* **2007**; 68(3): 636-45.
[152] Goodarzi H, Katanforoush A, Torabi N, Najafabadi HS. Solvent accessibility, residue charge and residue volume, the three ingredients of a robust amino acid substitution matrix. *J Theor Biol* **2007**; 245(4): 715-25.
[153] Huang LT, Saraboji K, Ho SY, Hwang SF, Ponnuswamy MN, Gromiha MM. Prediction of protein mutant stability using classification and regression tool. *Biophys Chem* **2007**; 125(2-3): 462-70.
[154] Parthiban V, Gromiha MM, Hoppe C, Schomburg D. Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins* **2007**; 66(1): 41-52.
[155] Li MH, Lin L, Wang XL, Liu T. Protein-protein interaction site prediction based on conditional random fields. *Bioinformatics* **2007**; 23(5): 597-604.
[156] Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci* **2006**; 15(5): 1017-29.
[157] Nguyen C, Gardiner KJ, Cios K. A hidden markov model for predicting protein interfaces. *J Bioinf Comp Biol* **2007**; 5(3): 739-53.
[158] Cheng J, Sweredoski M, Baldi P. DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Mining Knowl Disc* **2006**; 13(1): 1-10.
[159] Cao B, Porollo A, Adamczak R, Jarrell M, Meller J. Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. *Bioinformatics* **2006**; 22(3): 303-9.
[160] Punta M, Rost B. PROFcon: novel prediction of long-range contacts. *Bioinformatics* **2005**; 21(13): 2960-8.
[161] Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* **2007**; 8: 113.
[162] Chou KC, Chen NY. The biological functions of low-frequency phonons. *Sci Sin* **1977**; 20: 447-57.
[163] Chou KC. Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys Chem* **1988**; 30: 3-48.
[164] Pedersen TG, Sigurskjold BW, Andersen KV, *et al*. A nuclear-magnetic-resonance study of the hydrogen-exchange behavior of lysozyme in crystals and solution. *J Mol Biol* **1991**; 218: 413-26.
[165] Atilgan AR, Akan P, Baysal C. Small-World Communication of Residues and Significance for Protein Dynamics. *Biophys J* **2004**; 86: 85-91.
[166] Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol* **2003**; 330: 719-34.
[167] Kitchen J, Saunders RE, Warwicker J. Charge environments around phosphorylation sites in proteins. *BMC Struct Biol* **2008**; 8: 19.
[168] Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **2004**; 55: 1005-13.
[169] Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* **2005**; 58(2): 321-8.
[170] Nishikawa K, Matsuo Y. Development of pseudoenergy potentials for assessing protein 3D-1D compatibility and detecting weak homologies. *Protein Eng* **1993**; 6: 811-20.
[171] Saitoh S, Nakai T, Nishikawa K. Geometrical constraint approach for reproducing the native backbone conformation of a protein. *Proteins* **1993**; 15: 191-204.
[172] Ota M, Nishikawa K. Assessment of pseudo-energy potentials by the best-five test: a new use of the three-dimensional profiles of proteins. *Protein Eng* **1997**; 10: 339-51.
[173] Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins* **2005**; 59: 38-48.
[174] Ishida T, Nakamura S, Shimizu K. Potential for assessing quality of protein structure based on contact number prediction. *Proteins* **2006**; 64: 940-7.
[175] Nakai T, Kidera A, Nakamura H. Intrinsic nature of the three-dimensional structure of proteins as determined by distance geometry with good sampling properties. *J Biomol NMR* **1993**; 3: 19-40.
[176] Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. *Nature* **2000**; 409: 641-5.

[177] Huang YM, Bystroff C. Improved pairwise alignments of proteins in the twilight zone using local structure predictions. *Bioinformatics* **2006**; 22: 413-22.

[178] Kurgan L, Rahbari M, Homaeian L. Impact of the predicted protein structural content on prediction of structural classes for the twilight zone proteins. *Proceedings of 5th International Conference on Machine Learning and Applications* **2006**; 180-186.

[179] Smith J, Diez G, Klemm AH, Schewkunow V, Goldmann WH. CapZ-lipid membrane interactions: A computer analysis. *Theor Biol Med Model* **2006**; 3: 33-7.

[180] Concepcion GP, David MP, Padlan EA. Why don't humans get scrapie from eating sheep? A possible explanation based on secondary structure predictions. *Med Hypotheses* **2005**; 64: 919-24.

[181] Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet* **2006**; 2: 529-36.

[182] Gong H, Isom DG, Srinivasan R, Rose GD. Local secondary structure content predicts folding rates for simple, two-state proteins. *J Mol Biol* **2003**; 327: 1149-54.

[183] Huang JT, Cheng JP. Prediction of folding transition-state position (T) of small, two-state proteins from local secondary structure content. *Proteins* **2007**; 68: 218-22.

[184] Dobson PD, Doig AJ. Predicting enzyme class from protein structure without alignments. *J Mol Biol* **2005**; 345: 187-99.

[185] Dobson PD, Doig AJ. Distinguishing enzyme structures from non-enzymes without alignments. *J Mol Biol* **2003**; 330: 771-83.

Received: June 11, 2008 Revised: July 24, 2008 Accepted: July 29, 2008