Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4628025, 7 pages
http://dx.doi.org/10.1155/2016/4628025

Research Article

Characterization and Prediction of Protein Flexibility Based on Structural Alphabets

Qiwen Dong,[1] Kai Wang,[2] Bin Liu,[3] and Xuan Liu[4]

[1] *Institute for Data Science and Engineering, East China Normal University, Shanghai 200062, China*
[2] *College of Animal Science and Technology, Jilin Agricultural University, Changchun 130118, China*
[3] *School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China*
[4] *College of Engineering, Shanghai Ocean University, Shanghai 201303, China*

Correspondence should be addressed to Xuan Liu; xliu@shou.edu.cn

Received 13 May 2016; Accepted 2 August 2016

Academic Editor: Xun Lan

Motivation. To assist efforts in determining and exploring the functional properties of proteins, it is desirable to characterize and predict protein flexibilities. *Results.* In this study, the conformational entropy is used as an indicator of the protein flexibility. We first explore whether the conformational change can capture the protein flexibility. The well-defined decoy structures are converted into one-dimensional series of letters from a structural alphabet. Four different structure alphabets, including the secondary structure in 3-class and 8-class, the PB structure alphabet (16-letter), and the DW structure alphabet (28-letter), are investigated. The conformational entropy is then calculated from the structure alphabet letters. Some of the proteins show high correlation between the conformation entropy and the protein flexibility. We then predict the protein flexibility from basic amino acid sequence. The local structures are predicted by the dual-layer model and the conformational entropy of the predicted class distribution is then calculated. The results show that the conformational entropy is a good indicator of the protein flexibility, but false positives remain a problem. The DW structure alphabet performs the best, which means that more subtle local structures can be captured by large number of structure alphabet letters. Overall this study provides a simple and efficient method for the characterization and prediction of the protein flexibility.

1. Introduction

Proteins are dynamic molecules that are in constant motion. Their conformations are depending on environmental factors like temperature, pH, and interactions [1]. Some regions are more susceptible to change than others. Such motions play a critical role in many biological processes, such as protein-ligand binding [2], virtual screening [3], antigen-antibody interactions [4], protein-DNA binding [5], structure-based drug discovery [6], and fold recognition [7, 8].

Many studies try to predict protein flexibilities using either sequence or structure information of proteins [9]. Sonavane et al. [10] analyzed the local sequence features and the distribution of *B*-factor in different regions of protein three-dimensional structures. Yuan et al. [11] adopted support vector regression (SVR) approach with multiple sequence alignment as input to predict the *B*-factor distribution of a protein from its sequence. Schlessinger and Rost [12] found that flexible residues differ from regular and rigid residues in local features such as secondary structure, solvent accessibility, and amino acid preferences. They combined these local features and global evolution information for protein flexibility prediction. Several sequence-based *B*-factor prediction methods were compared by Radivojac et al. [13]. Different models have been proposed to predict *B*-factor distribution based on protein atomic coordinates. The normal mode analysis can identify the most mobile parts of the protein as well as their directions by focusing on a few Cα atoms that move the most [14, 15]. The translation liberation screw model [16] simplified the protein as a rigid body with movement along translation, liberation, and screw axes. The Gaussian network model (GNM) [17] transformed a protein as an elastic network of Cα atoms that fluctuate around their mean positions. Recently, Yang et al. [18] predicted the *B*-factor by combining local structure assembly variations with sequence-based and structure-based profiling. There are also many other methods for protein flexibility prediction [19–21].

All the above methods use the *B*- or temperature factors produced by X-ray crystallography to elucidate flexibilities of proteins. The *B*-factor reflects the degree of thermal motion and static disorder of an atom in a protein crystal structure [22]. However, there is noise in experimentally determinate *B*-factor. Many factors can affect the value of *B*-factor such as the overall resolution of the structure, crystal contacts, and, importantly, the particular refinement procedures [23]. *B*-values from different structures can therefore not be reasonably compared [12]. Some researchers considered that the upper limit of accuracy for the prediction of *B*-factors is no more than 80% [11].

Protein structures are not static and rigid. The polypeptide backbones and especially the side chains are constantly moving due to thermal motion and the kinetic energy of the atoms (Brownian motion) [24]. Recent study [1] used the continuum prediction of secondary structures to identify the region undergoing conformational change. Other researchers have pointed out that continuous secondary structure assignment can capture protein flexibility [25]. Furthermore, the MolMovDB database [26] consists of structures that are experimentally determinate to exhibit conformational flexibility enabling a variety of protein motions. The Morph Server [27] in particular has been used by many scientists to analyze pairs of conformations and produce realistic animations.

The present work aims to explore whether the predicted conformations from the protein sequences can characterize their flexibilities or not. To achieve this goal, a simplified description of protein structure has to be provided first. The protein secondary structure offers only a summary of general backbone conformation and of local interactions through hydrogen bonding. The DSSP program [28] provides 8-class secondary structures. However, most secondary structures prediction methods only predict 3-class states with nearly 80% accuracy [29, 30]. The secondary structures are very crude description of protein backbone structures. Recently, many studies try to describe protein structures in a more refined manner. Toward this goal, many fragment libraries or structure alphabets (SA) have been presented either in Cartesian coordinates space or in torsion angles space [31–33]. Camproux et al. first derived a 12-letter alphabet of fragments by Hidden Markov Model [34] and then extended to 27 letters by Bayesian information criterion [35]. De Brevern et al. [36] proposed a 16-letter alphabet generated by a self-organizing map based on a dihedral angle similarity measure. The prediction accuracy of local three-dimensional structure has been steadily increased by taking sequence information and secondary structure information into consideration [37]. A comprehensive evaluation of these and other structural alphabets is performed by Karchin et al. [38].

In this study, we first explore whether the conformation variants can capture protein flexibility. The multiple conformations of proteins are taken from the Baker decoy sets [39]. Each three-dimensional conformation is represented by the one-dimensional series of letters from a structural alphabet. Four different structure alphabets, including the secondary structure in 3-class and 8-class, the PB structure alphabet [37], and the DW structure alphabet [40], are investigated here. Here, the conformational entropy is used to quantitatively indicate the flexibility. The results show that the conformational entropy has high correlation with *B*-factor. We then predict the protein flexibility from basic amino acid sequence. The structure alphabet letters of proteins are predicted using only sequence information and the entropy function of the predicted class distribution is used to be indicators of protein flexibilities. Experiment is performed on a subset of the MolMovDB database [26]. The results indicate that the conformational entropy is a good indicator of protein flexibility.

2. Materials and Method

2.1. Dataset. Three datasets are used in this study for different experimental validation.

The first dataset is taken from the work of Bodén and Bailey [1], which is used for the prediction of protein flexibility. This dataset contains 171 nonredundant protein sequences, in which no pair of sequences has larger than 20% sequence identity. All the proteins exhibit conformational flexibility according to the comprehensive database of macromolecular movements (MolMovDB) [26]. Each sequence in this dataset has been annotated with a list of residue positions that have more than one local structure according to the structure alphabets.

The second dataset is used to train the support vector machine which is used for the local structure predictions of proteins. This dataset is a subset of PDB database [41] obtained from the PISCES [42] web-server. There is less than 25% sequence identity between any two proteins and any protein has a resolution better than 2.5 Å. The structures with missing atoms and chain breaks are excluded. The proteins that show homologue with the proteins from the first dataset are also excluded. The resulting dataset contains 928 protein chains.

The third dataset is used to test whether the changes of local structures can characterize the protein flexibility. To achieve this goal, a variant of conformations for one protein must be provided. We use the Baker decoy sets [39] previously used for the evaluation of knowledge-based mean force potentials. This dataset consists of 41 single domain proteins with varying degrees of secondary structures and lengths from 25 to 87 residues. Each protein is attached with about 1400 decoy structures generated by ab initio protein structure prediction method of Rosetta [43].

2.2. Training and Test of Local Structures. Many methods have been presented for the prediction of protein local structures. The dual-layer model has been adopted here, which is developed in our previous studies [44]. The method is based on the observation that neighboring local structures are strongly correlated. A dual-layer model is then designed for protein local structure prediction. The position specific score matrix (PSSM), generated by PSI-BLAST [45], is inputted to the first-layer classifier, whose output is further enhanced by a second-layer classifier. At each layer, a variant of classifiers can be used, such as support vector machine (SVM) [33], neural network (NN) [46], Hidden Markov Models (HMM). In this study, the SVM is selected as the classifier, since its performance is better than those of other classifiers. Experimental results show that the dual-layer model provides an efficient method for protein local structure prediction.

2.3. Characterization of Protein Flexibilities by Conformational Changes. The conformations of proteins are represented by the local structures in the form of a structural alphabet. All the local structure types can be referred to as structure alphabet. Four different structure alphabets, including the secondary structure in 3-class and 8-class, the PB structure alphabet [37], and the DW structure alphabet [40], are investigated here. The three-dimensional protein structures can be represented by one-dimensional structure alphabet sequences according to a specific structure alphabet. Given a protein and its variable conformations, we can convert them into several structure alphabet sequences. The changes of local structures can be used to characterize the protein flexibility. For example, there is a protein sequence $a_1, a_2, \ldots, a_n$. Its three-dimensional structures and conformations are labeled as structure alphabet sequences; we then obtained a structure alphabet matrix $a_{11}, a_{12}, \ldots, a_{nm}$, where a_{ij} is the probability of the structure alphabet letter of the jth conformation at the amino acid position i, n is the length of the protein sequence, and m is the total number of letters in the structure alphabet. The conformational entropy is then used as an indicator of the protein flexibility:

$$H(i) = -\sum_{j=1}^{m} a_{ij} \ln a_{ij}, \tag{1}$$

where $H(i)$ is the conformational entropy of the protein at sequence position i.

The correlation between the conformational entropies and the B-factors is calculated as follows:

$$\mathrm{cc} = \frac{\sum_{i=1}^{n} (H_i - \mathrm{Ave}(H)) (B_i - \mathrm{Ave}(B))}{\sqrt{\left[\sum_{i=1}^{n} (H_i - \mathrm{Ave}(H))^2\right] \left[\sum_{i=1}^{n} (B_i - \mathrm{Ave}(B))^2\right]}}, \tag{2}$$

where B_i is the B-factor of the protein at sequence position i and Ave(H) and Ave(B) are the average of the conformational entropy and the average of B-factor of the protein.

2.4. Prediction of Protein Flexibilities by Local Structure Entropies. Let the predicted local structure for a given residue be $Y = Y_1, \ldots, Y_m$, where Y_j is the probability that the residue is in the jth local structure class, and m is the number of local structure classes: 3 for 3-class secondary structure alphabet, 8 for 8-class secondary structure alphabet, 16 for PB structure alphabet, and 28 for DW structure alphabet. The conformation entropy of a residue is defined as

$$H = -\sum_{j=1}^{m} Y_j \ln Y_j. \tag{3}$$

High entropy indicates relative disorder. Low entropy indicates relative order.

2.5. Performance Metrics. The following measures are used to evaluate the prediction of protein flexibilities: sensitivity, specificity, precision, and the Receiver Operator Characteristic (ROC) curves, which are defined as follows:

$$\begin{aligned} \text{Sensitivity} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \text{Specificity} &= \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \\ \text{Precision} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \end{aligned} \tag{4}$$

where TP is the number of true positives (flexible residues correctly classified as flexible residues), FP is the number of false positives (rigid residues incorrectly classified as flexible residues), TN is the number of true negatives (rigid residues correctly classified as rigid residues), and FN is the number of false negatives (flexible residues incorrectly classified as rigid residues).

The ROC curve is plotted with true positives as a function of false positives for varying classification thresholds. A ROC score is the normalized area under the ROC curve. A score of 1 indicates the perfect separation of positive samples from negative samples, whereas a score of 0 denotes that none of the sequences selected by the algorithm is positive.

3. Results and Discussions

3.1. Local Structure Prediction. Four different structure alphabets are used in this study. They are the secondary structure in 3-class and 8-class, the PB structure alphabet [37], and the DW structure alphabet [40]. All of them are the description of the local structures of proteins.

The 3-class secondary structure provides a three-state description of backbone structures: helices, strands, and coils. The 8-class secondary structure provides a more detail description [28]. However, this description of protein structures is still very crude [47].

Two other structure alphabets are investigated in this study: the DW structure alphabet and the PB structure alphabet. They are represented in Cartesian coordinate space and in torsion angles space, respectively. The PB alphabet [37] is composed of 16 prototypes, each of which is 5-residue in length and represented by 8 dihedral angles. This structure alphabet remains valid although the size of the databank becomes large [48]. The DW structure alphabet is developed in our previous study [40], which is represented in Cartesian coordinate space. This structure alphabet contains 28 prototypes with lengths of 7 residues.

The dual-layer model is used to predict the local structures of proteins [44]. The experiment is performed on the second dataset. The Q-score is used to assess the prediction results, that is, the proportion of structure alphabet prototypes correctly predicted. This score is equivalent to the Q_3 value for secondary structure prediction. After 5-fold cross-validation, the results are shown in Table 1. The accuracy of secondary structure prediction is comparable with the currently state-of-the-art method [29], while the performances of the other two structure alphabets are significantly better

Table 1: The average Q-scores of local structure prediction for the four structure alphabets.

	Sec3	Sec8	PB	DW
Number of letters	3	8	16	28
Single-layer model	0.756	0.593	0.564	0.432
Dual-layer model	0.765	0.614	0.585	0.456

The single-layer model uses the position specific score matrix (PSSM) as input and output probability of the structure alphabet letters. The dual-layer model adds an additional classifier, which uses the output of single-layer model as input and output final prediction. For both models, the support vector machine is used as the classifiers.

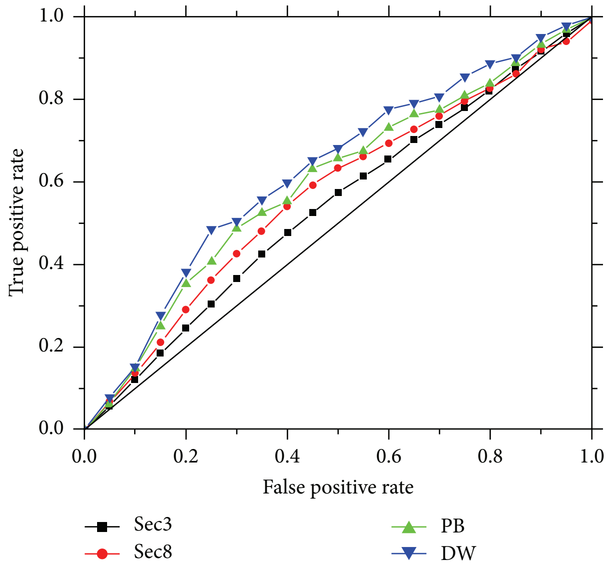

Figure 1: The ROC curve of the proposed method by using different structure alphabets on the set of 171 protein sequences.

than those of other related works [33, 37, 49, 50]. For detailed results, please refer to Dong et al. [44].

3.2. Results for the Characterization of Protein Flexibilities. Since proteins are dynamic molecules, we can investigate whether the conformational changes can capture protein flexibilities. The protein structures are represented by structure alphabet sequences. The conformational entropy is used as an indicator of protein flexibility. The experiment is performed on the third dataset.

The initial results demonstrate that some of the proteins show high correlations between the conformational entropies and the B-factors while the other proteins show low and even negative correlations. After detail analysis, we find that the correlations are influenced by the distribution of the decoy structures. Uniform distribution often leads to high correlation. The decoy structures are first classified by the Root-Mean-Squared Deviation (RMSD) with the native structures. We then select the decoy structures so that they are approximately uniform distribution between different classes. Some of the proteins and the correlations and are listed at Table 2 together with the number of decoy structures. As the number of letters increases, the correlations also increase.

According to the law of thermodynamics, the native structure is the one that has the lowest energy. Since proteins are dynamically molecular in living organisms, their structures often fluctuate around the native state. The decoy sets used here are generated by the well-known Rosetta algorithm [43]. These sets contain many decoy structures whose energies are close to the native one. The conformational entropies are then derived from the decoy sets. Some of the conformational entropies show high correlation with the protein flexibilities. However, the decoy sets are not the true stories; there still are some proteins that show low correlations between the entropies and the B-factors (data not shown). This experiment only tries to investigate whether the conformational changes can capture protein flexibilities. If the true decoy sets can be obtained, we can give a definite answer. However, obtaining the true decoy sets is costly and labor-intensive work.

3.3. Results for the Prediction of Protein Flexibilities. The experiment is performed on the first dataset. Each residue is labeled as a rigid or flexible residue. The animations of protein motions provided by the MolMovDB database [26] are converted into structure alphabet letter sequences by the specific structure alphabet. If a residue changes its structure alphabet letter among the animations, it is labeled as flexible residue. Otherwise, it is labeled as rigid residue.

During the prediction process, the protein local structures are first predicted from amino acid sequence by the dual-layer model, and then the entropy function is applied to the predicted class distribution for each residue. Residues with entropy larger than a given threshold T are predicted to be flexible residues. Otherwise, they are predicted to be rigid residues. Following the work of Bodén and Bailey [1] we use the mean entropy of all residues in our conformation variability dataset as the threshold T.

The results of the four structure alphabets are shown in Table 3. The corresponding Receiver Operator Characteristic (ROC) curves are given at Figure 1. The different structure alphabets get different number of positive (flexible) and negative (rigid) samples. As the number of letters in the structure alphabet increases, the number of positive samples increases and the prediction performance also increases, which means that more subtle local structures can be captured by large number of structure alphabet letters. Particularly, the precision and ROC scores steadily increase. Overall the DW structure alphabet gets the best performance.

The results obtained here are similar to the work of Bodén and Bailey [1]. The precisions of this study are higher than that of Bodén and Bailey (0.05 for Sec3 and 0.12 for Sec8), but the ROC scores are a little lower than of Bodén and Bailey (0.61 for Sec3 and 0.64 for Sec8). The main differences of this study to that of Bodén and Bailey lie in two aspects. The first one is that the additional two structure alphabets (the PB and DW structure alphabet) are investigated here. The second one is that a decoy set is used to explore whether the conformation change can capture protein flexibility.

Table 2: The correlations between the conformational entropies and the B-factors.

ID	<3[a]	3-4	4-5	5-6	>6	Sec3[b]	Sec8	PB	DW
1res	73	73	73	7	4	0.1105	0.1505	0.2454	0.2605
1am3	571	177	162	161	400	0.1139	0.2993	0.4110	0.5149
1r69	389	119	284	228	300	0.2028	0.4040	0.3909	0.3794
1utg	1	20	401	290	300	0.2003	0.2990	0.2729	0.1653
1a32	364	125	95	142	300	0.2819	0.4818	0.5077	0.4145
1mzm	9	306	317	171	300	0.0118	0.2734	0.3353	0.3144
1hyp	1	0	34	270	300	0.1491	0.3579	0.1893	0.2889
1cei	1	0	4	64	300	0.0821	0.3583	0.4335	0.4932
1pgx	219	342	182	391	300	0.0264	0.2843	0.3339	0.3674
5icb	3	142	481	225	300	0.4255	0.5660	0.5433	0.5635
Ave	163.1	130.4	203.3	194.9	280.4	0.1604	0.3474	0.3663	0.3762

[a] Shown in the table are the numbers of decoy structures in this class.
[b] Shown in the table are the correlations measured by the specific structure alphabet.

Table 3: Prediction performance of the protein flexibilities by different structure alphabets.

SA[a]	No. po[b]	No. ne[c]	Sensitivity	Specificity	Precision	ROC
Sec3	6152	54737	0.6291	0.4543	0.1109	0.5457
Sec8	9468	51421	0.5887	0.5677	0.1942	0.5741
PB	10625	50264	0.6209	0.5521	0.2114	0.5901
DW	16012	44877	0.6399	0.5725	0.2586	0.6193

[a] The structure alphabet types.
[b] The number of positive samples (flexible residues).
[c] The number of negative samples (rigid residues).

4. Conclusion

In this study we provide a simple and efficient method for the characterization and prediction of the protein flexibility. We first validate that the conformational change can capture protein flexibility and then predict protein flexibility from primary sequences. The results show that conformational entropy is a good indicator of protein flexibility. Four structure alphabets with different number of letters are investigated. Future work will aim at exploring other structure alphabets that can provide detail description of protein backbone structures and even the side-chain structures.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this article.

References

[1] M. Bodén and T. L. Bailey, “Identifying sequence regions undergoing conformational change via predicted continuum secondary structure,” *Bioinformatics*, vol. 22, no. 15, pp. 1809–1814, 2006.

[2] J. Li, J. Cai, H. Su et al., “Effects of protein flexibility and active site water molecules on the prediction of sites of metabolism for cytochrome P450 2C19 substrates,” *Molecular BioSystems*, vol. 12, no. 3, pp. 868–878, 2016.

[3] P. Manoharan, K. Chennoju, and N. Ghoshal, “Target specific proteochemometric model development for BACE1—protein flexibility and structural water are critical in virtual screening,” *Molecular BioSystems*, vol. 11, no. 7, pp. 1955–1972, 2015.

[4] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović, “Intrinsic disorder and protein function,” *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, 2002.

[5] H. J. Dyson and P. E. Wright, “Coupling of folding and binding for unstructured proteins,” *Current Opinion in Structural Biology*, vol. 12, no. 1, pp. 54–60, 2002.

[6] D. A. Antunes, D. Devaurs, and L. E. Kavraki, “Understanding the challenges of protein flexibility in drug design,” *Expert Opinion on Drug Discovery*, vol. 10, no. 12, pp. 1301–1313, 2015.

[7] Z. Feng and X. Hu, “Recognition of 27-class protein folds by adding the interaction of segments and motif information,” *BioMed Research International*, vol. 2014, Article ID 262850, 9 pages, 2014.

[8] J. Chen, B. Liu, and D. Huang, “Protein remote homology detection based on an ensemble learning approach,” *BioMed Research International*, vol. 2016, Article ID 5813645, 11 pages, 2016.

[9] A. Petrovich, A. Borne, V. N. Uversky, and B. Xue, “Identifying similar patterns of structural flexibility in proteins by disorder prediction and dynamic programming,” *International Journal of Molecular Sciences*, vol. 16, no. 6, pp. 13829–13849, 2015.

[10] S. Sonavane, A. A. Jaybhaye, and A. G. Jadhav, “Prediction of temperature factors from protein sequence,” *Bioinformation*, vol. 9, no. 3, pp. 134–140, 2013.

[11] Z. Yuan, T. L. Bailey, and R. D. Teasdale, "Prediction of protein B-factor profiles," *Proteins: Structure, Function, and Bioinformatics*, vol. 58, no. 4, pp. 905–912, 2005.

[12] A. Schlessinger and B. Rost, "Protein flexibility and rigidity predicted from sequence," *Proteins: Structure, Function and Genetics*, vol. 61, no. 1, pp. 115–126, 2005.

[13] P. Radivojac, Z. Obradovic, D. K. Smith et al., "Protein flexibility and intrinsic disorder," *Protein Science*, vol. 13, no. 1, pp. 71–80, 2004.

[14] V. Alexandrov, U. Lehnert, N. Echols, D. Milburn, D. Engelman, and M. Gerstein, "Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool," *Protein Science*, vol. 14, no. 3, pp. 633–643, 2005.

[15] W. G. Krebs, V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein, "Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic," *Proteins: Structure, Function and Genetics*, vol. 48, no. 4, pp. 682–695, 2002.

[16] J. Kuriyan and W. I. Weis, "Rigid protein motion as a model for crystallographic temperature factors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 7, pp. 2773–2777, 1991.

[17] T. Haliloglu, I. Bahar, and B. Erman, "Gaussian dynamics of folded proteins," *Physical Review Letters*, vol. 79, article 3090, 1997.

[18] J. Yang, Y. Wang, and Y. Zhang, "ResQ: an approach to unified estimation of *B*-factor and residue-specific error in protein structure prediction," *Journal of Molecular Biology*, vol. 428, no. 4, pp. 693–701, 2016.

[19] J. A. Kovacs, P. Chacón, and R. Abagyan, "Predictions of protein flexibility: first-order measures," *Proteins: Structure, Function and Genetics*, vol. 56, no. 4, pp. 661–668, 2004.

[20] K. Xia and G. W. Wei, "Stochastic model for protein flexibility analysis," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 88, no. 6, Article ID 062709, 2013.

[21] Y. Gu, D.-W. Li, and R. Brüschweiler, "Decoding the mobility and time scales of protein loops," *Journal of Chemical Theory and Computation*, vol. 11, no. 3, pp. 1308–1314, 2015.

[22] J. Drenth, *Principles of Protein Crystallography*, Springer, New York, NY, USA, 1994.

[23] D. E. Tronrud, "Knowledge-based B-factor restraints for the refinement of proteins," *Journal of Applied Crystallography*, vol. 29, no. 2, pp. 100–104, 1996.

[24] A. Sharma and E. S. Manolakos, "Efficient multicriteria protein structure comparison on modern processor architectures," *BioMed Research International*, vol. 2015, Article ID 563674, 13 pages, 2015.

[25] C. A. F. Andersen, A. G. Palmer, S. Brunak, and B. Rost, "Continuum secondary structure captures protein flexibility," *Structure*, vol. 10, no. 2, pp. 175–184, 2002.

[26] S. Flores, N. Echols, D. Milburn et al., "The Database of Macromolecular Motions: new features added at the decade mark," *Nucleic Acids Research*, vol. 34, pp. D296–D301, 2006.

[27] W. G. Krebs and M. Gerstein, "The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework," *Nucleic Acids Research*, vol. 28, no. 8, pp. 1665–1675, 2000.

[28] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[29] O. Dor and Y. Zhou, "Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 4, pp. 838–845, 2007.

[30] 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: Structure, Function, and Bioinformatics*, vol. 66, no. 3, pp. 664–670, 2007.

[31] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt, "Small libraries of protein fragments model native protein structures accurately," *Journal of Molecular Biology*, vol. 323, no. 2, pp. 297–307, 2002.

[32] J. B. Holmes and J. Tsai, "Some fundamental aspects of building protein structures from fragment libraries," *Protein Science*, vol. 13, no. 6, pp. 1636–1650, 2004.

[33] O. Sander, I. Sommer, and T. Lengauer, "Local protein structure prediction using discriminative models," *BMC Bioinformatics*, vol. 7, article 14, 2006.

[34] A. C. Camproux, P. Tuffery, J. P. Chevrolat, J. F. Boisvieux, and S. Hazout, "Hidden Markov model approach for identifying the modular framework of the protein backbone," *Protein Engineering*, vol. 12, no. 12, pp. 1063–1073, 1999.

[35] A. C. Camproux, R. Gautier, and P. Tufféry, "A hidden Markov model derived structural alphabet for proteins," *Journal of Molecular Biology*, vol. 339, no. 3, pp. 591–605, 2004.

[36] A. G. De Brevern, C. Etchebest, and S. Hazout, "Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks," *Proteins: Structure, Function and Genetics*, vol. 41, no. 3, pp. 271–287, 2000.

[37] C. Etchebest, C. Benros, S. Hazout, and A. G. De Brevern, "A structural alphabet for local protein structures: improved prediction methods," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, no. 4, pp. 810–827, 2005.

[38] R. Karchin, M. Cline, and K. Karplus, "Evaluation of local structure alphabets based on residue burial," *Proteins: Structure, Function and Genetics*, vol. 55, no. 3, pp. 508–518, 2004.

[39] J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker, "An improved protein decoy set for testing energy functions for protein structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 53, no. 1, pp. 76–87, 2003.

[40] Q.-W. Dong, X.-L. Wang, and L. Lin, "Methods for optimizing the structure alphabet sequences of proteins," *Computers in Biology and Medicine*, vol. 37, no. 11, pp. 1610–1616, 2007.

[41] A. Kouranov, L. Xie, J. de la Cruz et al., "The RCSB PDB information portal for structural genomics," *Nucleic Acids Research*, vol. 34, pp. D302–D305, 2006.

[42] G. Wang and R. L. Dunbrack Jr., "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[43] K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker, "Ab initio protein structure prediction of CASP III targets using ROSETTA," *Proteins*, vol. 37, no. 3, pp. 171–176, 1999.

[44] Q. Dong, X. Wang, L. Lin, and Y. Wang, "Analysis and prediction of protein local structure based on structure alphabets," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, no. 1, pp. 163–172, 2008.

[45] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[46] J. Hawkins and M. Bodén, "The applicability of recurrent neural networks for biological sequence analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 3, pp. 243–253, 2005.

[47] W. Chu, Z. Ghahramani, A. Podtelezhnikov, and D. L. Wild, "Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 3, no. 2, pp. 98–113, 2006.

[48] A. G. de Brevern, "New assessment of a structural alphabet," *In Silico Biology*, vol. 5, no. 3, pp. 283–289, 2005.

[49] C. Benros, A. G. De Brevern, C. Etchebest, and S. Hazout, "Assessing a novel approach for predicting local 3D protein structures from sequence," *Proteins: Structure, Function and Genetics*, vol. 62, no. 4, pp. 865–880, 2006.

[50] T. Tang, J. Xu, and M. Li, "Discovering sequence-structure motifs from protein segments and two applications," *Pacific Symposium on Biocomputing*, pp. 370–381, 2005.