COMPREHENSIVE ANALYSIS OF SEQUENCE-STRUCTURE RELATIONSHIPS IN THE LOOP REGIONS OF PROTEINS

SHUGO NAKAMURA[1] KENTARO SHIMIZU[1]
shugo@bi.a.u-tokyo.ac.jp shimizu@bi.a.u-tokyo.ac.jp

[1] *Department of Biotechnology, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan*

Local sequence-structure relationships in the loop regions of proteins were comprehensively estimated using simple prediction tools based on support vector regression (SVR). End-to-end distance was selected as a rough structural property of fragments, and the end-to-end distances of an enormous number of loop fragments from a wide variety of protein folds were directly predicted from sequence information by using SVR. We found that our method was more accurate than random prediction for predicting the structure of fragments comprising 5, 9, and 17 amino acids; moreover, the extended loop fragments could be successfully distinguished from turn structures on the basis of their sequences, which implies that the sequence-structure relationships were significant for loop fragments with a wide range of end-to-end distances. These results suggest that many loop regions as well as helices and strands restrict the conformational space of the entire tertiary structure of proteins to some extent; moreover, our findings throw light on the mechanism of protein folding and prediction of the tertiary structure of proteins without using structural templates.

Keywords: protein; loop; end-to-end distance; prediction; sequence-structure relationship; support vector regression.

1. Introduction

The tertiary structure of a protein is related to its functional properties to a large extent and is determined by its entire amino acid sequence [3]. The structure of helices and strands, which are the major secondary structures of proteins, is related to their amino acid sequences; hence, their structure can be predicted on the basis of the local amino acid sequence. Recently, it was reported that the prediction accuracy for 3-state secondary structures (helices, strands, and other structures) is nearly 80% [1]. While variations in the tertiary structures of helices and strands are small, the tertiary structures of protein loops that connect helices and strands are thought to be conformationally flexible and largely affected by the surrounding residues.

However, researchers have reported that in some cases, local sequences determine local structures, including loop regions. Long-range interactions are essentially important for protein folding, but if local sequences determine local structures to some extent, it leads to restriction of conformational space of the entire

tertiary structure, greatly affecting its folding pathway. The sequence-structure relationships for turns in the loop regions have been analyzed [12, 18], and tools for predicting the existence of turns on the basis of amino acid sequences have been developed [10, 28, 29]. On the basis of clustering of amino acid sequences, Han *et al.* analyzed recurring local sequence motifs that cross protein-family boundaries [13]. In addition, they analyzed conservation of secondary and tertiary structures of sequence clusters [14, 15]. Bystroff *et al.* obtained "I-sites library," which is a group of fragments that have similar tertiary structures within sequence clusters [6], and applied the library to predict the tertiary structure of proteins on the basis of their amino acid sequences [7]. In several studies, a number of local structural motifs or the so-called "structural alphabets" were determined by clustering the tertiary structures of fragments deposited in a protein structure database. The assignment of these alphabets to a query fragment was predicted using the local amino acid sequences as inputs [4, 9, 17, 25, 30]. The prediction accuracy of loop modeling without the use of the structures of regions flanking the loop regions has also been reported [22].

The abovementioned studies have shown that in certain loop structures, the local sequences and structures are highly related. However, most of these studies used clustering of sequence and/or structure spaces, and information on fragments that were not included in the clusters was neglected. The researchers reported the prediction accuracy for all fragments, including helices and strands; however, very few of them have reported prediction with comprehensive accuracy for the loop regions. Moreover, they provided little information on the extent of sequence-structure relationships in loop regions other than the turns, such as the percentage of structures that can be predicted on the basis of the sequence information of that region.

In this study, we have comprehensively estimated local sequence-structure relationships in the loop regions of proteins. We estimated the local sequence-structure relationships by analyzing whether local amino acid sequence information could be used to determine the structural properties of that region. We did not predict the assignment of structural alphabets. Instead, we directly predicted the structural information of each fragment. We hypothesized that if the accuracy of our prediction is better than that of random prediction, it implies that local sequences determine local structures to some extent, and that the sequence-structure relationship can be detected using our prediction tool. It is expected that the marginal sequence-structure relationships can be detected using rough structural representations. We adopted end-to-end distance as a rough structural property of fragments. The end-to-end distance of protein fragments is one of the parameters that best represent the structures of fragments [19] and is used for structure prediction [11].

We could have used any prediction tool for our method of prediction, but few local structure prediction tools that were previously reported are available. Therefore, we developed a simple prediction tool based on support vector regression (SVR). Our tool was applied to an enormous number of protein fragments comprising 5, 9, and 17 amino acids retrieved from a wide variety of protein folds in the non-

Table 1. SVR parameters γ and C used for predictions.

Fragment length	5	9	17
Input: Amino acids	-	0.1,2.0	-
Input: PSSM	0.1,1.0	0.05,5.0	0.05,10.0

redundant structure database, and the prediction accuracy of our tool for the determination of loop structures was compared to the expected accuracy of random predictions.

2. Materials and Methods

2.1. *Preparation of Datasets*

We selected fold representative domains from the Structural Classification of Proteins (SCOP) 1.71 [5, 23], from which fragments were extracted. Membrane proteins and protein structures determined by nuclear magnetic resonance (NMR) were excluded. From 702 domain structures, 168,066, 164,703, and 158,704 fragments comprising 5, 9, and 17 amino acids, respectively, were collected; the fragments were allowed to overlap.

2.2. *Predictions Using SVR*

To determine the sequence-structure relationships in the loop regions, we used a tool for predicting end-to-end distance on the basis of sequence information. Since few suitable tools that could be applied to an enormous number of fragment data on a standalone machine and that could allow flexible specification of user-defined inputs were available, we developed a simple prediction tool based on SVR [26].

Epsilon-SVR with radial basis function (RBF) kernel implemented in the LIBSVM 2.83 package [8] was used, and the parameters for SVR — γ for RBF kernel and cost parameter C — were determined by a grid search to obtain the maximum correlation coefficient of the predicted and actual end-to-end distances (defined as the distance between Cα atoms of the N-terminal residue and the C-terminal residue of a fragment) on the basis of the 5-fold cross-validation testing described below. We used the following data as inputs: (1) amino acid sequence (fragments with 9 amino acids only) and (2) position-specific scoring matrix (PSSM). The number of dimensions of an input vector is $20 \times w$, where w is the fragment length. PSSMs were obtained by PSI-BLAST [2] of up to 5 iterations applied to the non-redundant database (nr) with default parameters; whole-domain sequences were used as query fragments. The γ and C pairs obtained for predictions are shown in Table 1.

2.3. *Calculation of Prediction Accuracy*

We estimated the prediction accuracy by using 5-fold cross validation. For 5-fold cross validation, we divided the collected fragment set into 5 subsets. Four subsets

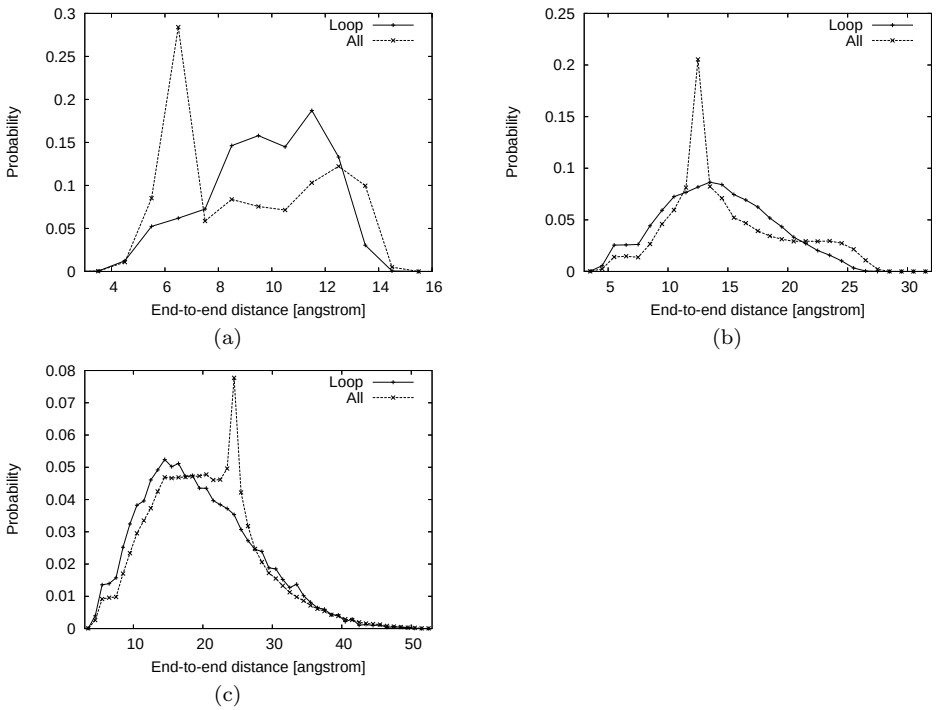

Fig. 1. End-to-end distance distributions for loop structures ("Loop") and fragments ("All") comprising (a) 5, (b) 9, and (c) 17 amino acids.

were used for the training of SVR, and the remaining subset was used for the test. To eliminate redundancy between the training and test data, we retrieved BLAST hits for proteins in the test data with an e-value threshold of < 10.0 and a database size of 100,000,000, using the proteins in the training data as queries.

Here, we define a loop fragment as a fragment in which more than half of the consecutive residues are coil residues (neither helix nor strand residues). For secondary structure assignment, we used DSSP [20]. The structure letters H, G, and I were assigned to helices; B and E were assigned to strands, and other structure letters were assigned to coils. To eliminate misassignment of beta strands in beta sheets between protein chains as coils, DSSP was applied to structural data of biological units obtained from the PQS server [16]. The total numbers of loop fragments were 57,752, 43,450, and 19,385 for fragments comprising 5, 9, and 17 amino acids, respectively.

Note that the training sets for our prediction include various fragments such as helices and strands, and not merely the sets of loop fragments. Figure 1(a)–(c) shows the end-to-end distance distributions of fragments comprising 5, 9, and 17 amino acids. The solid lines ("Loop") show the distributions of loop fragments, and the dashed lines ("All") show the distributions of all the fragments. Sharp peaks of distributions for all the fragments correspond to helices.

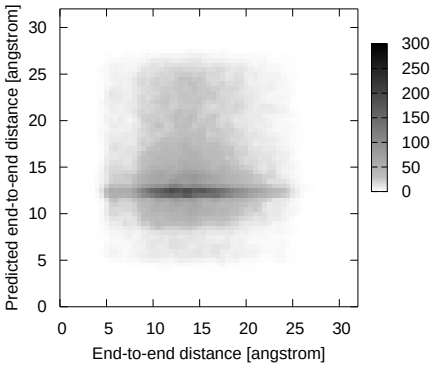

Fig. 2. An example of a histogram showing randomly predicted end-to-end distances against the actual end-to-end distances for loop fragments comprising 9 amino acids.

2.4. *Random Prediction*

The accuracy of random prediction was calculated in order to determine the significance of the accuracy of our prediction. As described above, our training sets included fragments with all kinds of structures such as helices, strands, and loops with end-to-end distance distributions that were very similar to the "All" lines in Fig. 1(a)–(c). Thus, the end-to-end distance for each fragment in the test set was randomly predicted on the basis of the end-to-end distributions of fragments in the training sets. A total of 1000 random predictions were made for each set comprising fragments with 5, 9, and 17 amino acids, and the average correlation coefficients were calculated to be 0.00015 ± 0.015, -0.00015 ± 0.015, and -0.00022 ± 0.022, respectively. Figure 2 shows, as an example, the random predictions for fragments comprising 9 amino acids.

2.5. *Dataset from CASP8 Targets*

To evaluate the dataset dependency of our results, we applied our prediction tool to fragments extracted from another set of protein structures. This set belonged to the "free modeling targets" of the eighth community-wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP8) [31]. Free modeling targets were proteins that did not have homologs with known tertiary structures when CASP8 was conducted. We used PSSMs calculated during the CASP8 prediction term as inputs in SVR. Thus, structural information on the homologs and the targets themselves, which were made public after CASP8, were not used for our prediction. From these targets, we extracted 864, 829, and 780 fragments comprising 5, 9, and 17 amino acids, respectively. Among the fragments comprising 5, 9, and 17 amino acids, 282, 177, and 88, respectively, were loop fragments. We applied our prediction tool to these fragments in the same manner as we did for the cross-validation testing described above.

Table 2. Correlation coefficients between the predicted and actual end–to-end distances for fragments comprising 5, 9, and 17 amino acids.

Fragment length	5		9		17	
	Loop	All	Loop	All	Loop	All
Input: Amino acids	-	-	0.313	0.434	-	-
Input: PSSM	0.485	0.673	0.502	0.666	0.446	0.551

3. Results and Discussion

Table 2 shows the correlation coefficients between the predicted and actual end-to-end distances for fragments comprising 5, 9, and 17 amino acids. The values for loop fragments are given in the "Loop" columns, and those for all the fragments in the "All" columns. In the case of random predictions, i.e., when there is no relationship between the local sequences and local structures, the correlation coefficients should almost be zero, as described in the random prediction section (2.4) and as shown in Fig. 2.

When amino acid sequence information was used as the input, the correlation coefficients were above 0.3 for loop fragments and above 0.4 for all the fragments comprising 9 amino acids; these values are better than those obtained in random prediction. When PSSM was used as the input, the prediction accuracies significantly improved; the correlation coefficients were above 0.4 for the loop fragments and above 0.5 for all the fragments comprising 5, 9, and 17 amino acids. Previous researches have reported that for backbone dihedral angle prediction, when PSSM was used as the input instead of amino acid sequences, the prediction accuracy improved by only a small extent [21]. We think that we were able to detect the effect of PSSM because we used end-to-end distance, which is a rough structural property, and not backbone dihedral angles, which are very sensitive to the tertiary structures of fragments.

Figure 3(a)–(c) are histograms showing the predicted and actual end-to-end distances for loop fragments comprising 5, 9, and 17 amino acids (PSSM input). In all the cases, there were obvious correlations between the predicted and actual end-to-end distance. For 53% of the fragments, the errors in the predicted end-to-end distances compared to the actual distances were below 20%. In particular, our simple prediction tool can accurately distinguish fragments with short end-to-end distances from those with long end-to-end distances. When the predicted end-to-end distance was greater than 20 Å for fragments comprising 9 amino acids, it was found that only 2.5% of them had an actual end-to-end distance of < 10 Å. These results suggest that the local structures of the extended loops and turns considerably depend on the local sequences. Figure 3(d) shows the receiver operating characteristic (ROC) curve for predicting whether the end-to-end distance of a fragment is longer than 20 Å for fragments comprising 9 amino acids. The area under the ROC curve (AUC) was 0.774, and the maximum Matthews correlation coefficient (MCC) was 0.292;

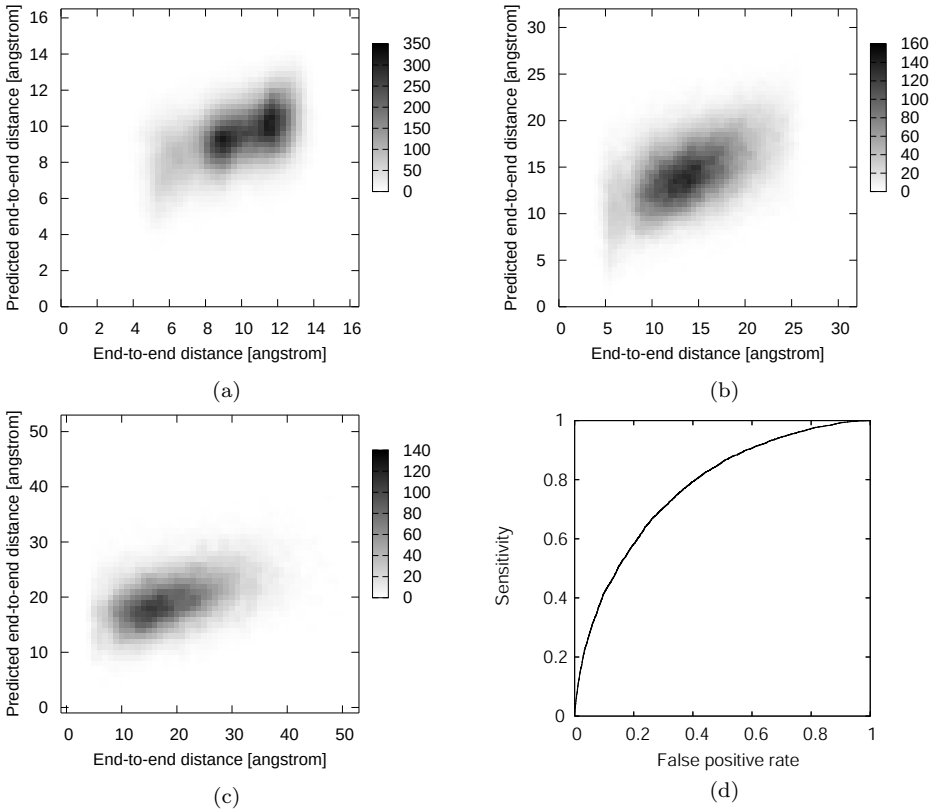

Fig. 3. Histograms depicting the predicted end-to-end distances against the actual end-to-end distances for loop fragments comprising (a) 5, (b) 9, and (c) 17 amino acids. (d) Receiver operating characteristic (ROC) curve for predicting whether the end-to-end distance of a fragment is longer than 20 Å for fragments comprising 9 amino acids.

these results show that our prediction method is better than random prediction for the prediction of extended loop structures.

Table 3 shows the correlation coefficients between the predicted and actual end-to-end distances for fragments comprising 5, 9, and 17 amino acids extracted from the CASP8 dataset. Though the numbers of data are small, the correlation coefficients for these fragments are comparable to those for the fragments in the 5-fold cross-validation test (Table 2), and the results show that the sequence-structure relationships detected by our prediction tool are generally independent of the datasets.

To elucidate the characteristics of amino acid sequences on the basis of the end-to-end distance of loop fragments, we analyzed the amino acid propensity for fragments comprising 9 amino acids. A similar tendency was observed for fragments comprising 5 and 17 amino acids. We divided the fragments into 4 classes on the basis of the end-to-end distances: < 10 Å, 10–15 Å, 15–20 Å, and > 20 Å. Subsequently, we calculated the amino acid propensity for fragments of each distance class

Table 3. Correlation coefficients between the predicted and actual end-to-end distances for fragments comprising 5, 9, and 17 amino acids in the CASP8 dataset.

Fragment length	5		9		17	
	Loop	All	Loop	All	Loop	All
Input: PSSM	0.472	0.631	0.525	0.625	0.240	0.518

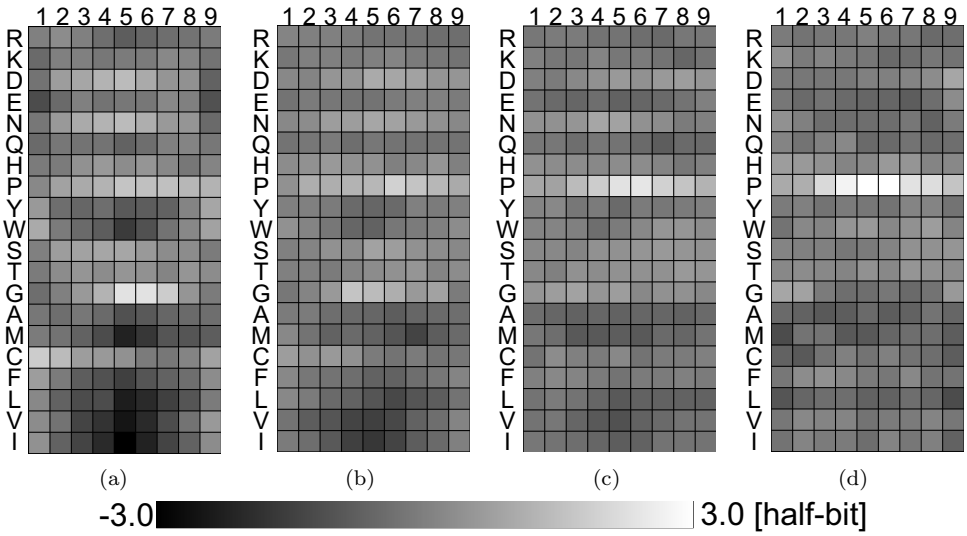

Fig. 4. Amino acid propensity at each position in the fragments comprising 9 amino acids of each distance class. (a) Propensities for fragments with end-to-end distances of < 10 Å, (b) 10–15 Å, (c) 15–20 Å, and (d) > 20 Å.

(Fig. 4(a)–(d)). The propensity P_{ijk} (in half-bit unit) of amino acid a_i at position j for class k was obtained by the following equation:

$$P_{ijk} = 2\log_2\left(\frac{n_j^k(a_i)/N_j^k}{n(a_i)/N}\right),$$

where $n_j^k(a_i)$ is the frequency of amino acid a_i at position j in the loop fragments of class k; N_j^k is the frequency of any amino acid at position j in the loop fragments of class k; $n(a_i)$ is the frequency of amino acid a_i at any position in all the fragments of any class; and N is the frequency of any amino acid at any position in all the fragments of any class.

The characteristics of amino acid propensity for fragments in each distance class were clearly observed. In the case of fragments with short end-to-end distances (Fig. 4(a)), disfavor for hydrophobic residues such as Met and Trp and preference for Asp, Asn, Pro, and Gly were obvious, as described by previous studies [12, 18]. This tendency gradually disappeared as the end-to-end distance of the fragment

Table 4. Distributions of the number of Pro residues in all, loop, and extended loop (end-to-end distance of > 20 Å) fragments comprising 9 amino acids.

Number of Pro residues	0	1	2	3	4	5
All	0.662	0.276	0.053	0.008	0.001	0.0001
Loop	0.484	0.381	0.111	0.021	0.003	0.0003
Loop (> 20 Å)	0.380	0.422	0.153	0.034	0.009	0.002

increased (Fig. 4(b,c)), and fragments with end-to-end distances of > 20 Å showed significant preference for Pro (Fig. 4(d)). Table 4 shows the distributions of the number of Pro residues in all, loop, and extended loop (end-to-end distance, > 20 Å) fragments comprising 9 amino acids. Among the extended loop fragments, fragments with more than 2 Pro residues were not frequently observed, but the average number of Pro residues per extended loop fragment (0.88) was more than double that per fragment in the case of all the fragments (0.41).

The extended loop fragments that show a preference for Pro residues correspond to the "Pro-rich extended fragments" proposed by Ikeda *et al.* [19]. These tendencies can also be observed for PSSM, and they may explain why SVR can be used to predict the end-to-end distance for extended fragments. Recently, the polyproline II-type (PPII) structure has received attention as a type of extended structure of polypeptides [9, 24], and a prediction tool for determining the PPII structure on the basis of amino acid sequences has been developed [27]. According to the descriptions of PPII residues reported earlier [27], only approximately 10% of fragments with an end-to-end distance of > 20 Å have at least 5 PPII residues. Our results suggest that the sequence-structure relationships in extended loop fragments are not restricted to PPII structures.

4. Conclusion

In this study, we used a simple machine-learning-based prediction tool to predict the end-to-end distances of loop fragments comprising 5, 9, and 17 amino acids. Our prediction was found to be more accurate than random prediction. Extended loop fragments could be well distinguished from fragments with short end-to-end distances. Our findings throw light on the mechanism of protein folding and the prediction of the tertiary structure of proteins without using structural templates.

Most of the local structure prediction tools that were recently proposed are based on the prediction of a sequence of structural alphabets. The prediction of structural alphabets is a kind of classification prediction and does not consider structural similarity among these alphabets; moreover, this type of prediction provides little information if the prediction fails. Instead, we directly predicted continuous values of the end-to-end distance of fragments as a rough structural property, and we think this is the reason we could detect significant sequence-structure relationships for fragments with a wide range of end-to-end distances.

It is important to note that our prediction tool is rather simple, and hence, the detected sequence-structure relationships may be lower limit. The use of our tool in combination with other local structure prediction tools such as alpha-, gamma-, and beta-turn prediction [10, 28, 29], or local structure prediction tools based on structural alphabets will enhance the ability to detect further relationships between local sequences and structures.

References

[1] Aloy, P., Stark, A., Hadley, C., and Russell, R.B., Predictions without templates: new folds, secondary structures, and contacts in CASP5, *Proteins*, 53:436–456, 2003.
[2] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25(17):3389–3402, 1997.
[3] Anfinsen, C.B., Principles that govern the folding of protein chains, *Science*, 181(4096):223–230, 1973.
[4] Benros, C., de Brevern, A.G., Etchebest, C., and Hazout, S., Assessing a novel approach for predicting local 3D protein structure from sequence, *Proteins*, 62:865–880, 2006.
[5] Brenner S.E., Koehl, P., and Levitt, M., The ASTRAL compendium for sequence and structure analysis, *Nucleic Acids Res.*, 28:254–256, 2000.
[6] Bystroff, C. and Baker, D., Prediction of local structure in proteins using a library of sequence-structure motifs, *J. Mol. Biol.*, 281(3):565–577, 1998.
[7] Bystroff, C., Thorsson, V., and Baker, D., HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins, *J. Mol. Biol.*, 301(1):173–190, 2000.
[8] Chang, C.C. and Lin, C.J., LIBSVM: a library for support vector machines, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
[9] De Brevern, A.G., Etchebest, C., Benros, C., and Hazout, S., "Pinning strategy": a novel approach for predicting the backbone structure in terms of protein blocks from sequence, *J. Biosci.*, 32(1):51–70, 2007.
[10] Fuchs, P.F.J. and Alix, A.J.P., High accuracy prediction of beta-turns and their types using propensities and multiple alignments, *Proteins*, 59:828–839, 2005.
[11] Gront, D. and Kolinski, A., A new approach to prediction of short-range conformational propensities in proteins, *Bioinformatics*, 21(7):981–987, 2005.
[12] Guruprasad, K. and Rajkumar, S., Beta- and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials, *J. Biosci.*, 25(2):143–156, 2000.
[13] Han, K.F. and Baker, D., Recurring local sequence motifs in proteins, *J. Mol. Biol.*, 251(1):176–187, 1995.
[14] Han, K.F. and Baker, D., Global properties of the mapping between local amino acid sequence and local structure in proteins., *Proc. Nat'l Acad. Sci. U. S. A.*, 93(12):5814–5818, 1996.
[15] Han, K.F., and Bystroff, C., and Baker, D., Three-dimensional structures and contexts associated with recurrent amino acid sequence patters, *Protein Sci.*, 6:1587–1590, 1997.
[16] Henrick, K. and Thornton, J.M., PQS: a protein quaternary structure file server, *Trends Biochem. Sci.*, 23(9):358–361, 1998.
[17] Hunter, C.G. and Subramaniam, S., Protein local structure prediction from sequence, *Proteins*, 50:572–579, 2003.

[18] Hutchinson, E.G. and Thornton, J.M., A revised set of potentials for beta-turn formation in proteins, *Protein Sci.*, 3:2207–2216, 1994.
[19] Ikeda, K., Tomii, K., Yokomizo, T., Mitomo, D., Maruyama, K., Suzuki, S., and Higo, J., *Protein Sci.*, 14:1253–1265, 2005.
[20] Kabsch, W. and Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22(12):2577–2637, 1983.
[21] Kuang, R., Leslie, C.S., and Yang, A.-S., Protein backbone angle prediction with machine learning approaches, *Bioinformatics*, 20(10):1612–1621, 2004.
[22] Mönnigmann, M. and Floudas, C.A., Protein loop structure prediction with flexible stem geometries, *Proteins*, 61:748–762, 2005.
[23] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol, Biol.*, 247:536–540, 1995.
[24] Rath A., Davidson, A.R., and Deber, C.M., The structure of "unstructured" regions in peptides and proteins: role of the polyproline II helix in protein folding and recognition, *Biopolymers*, 80:179–185, 2005.
[25] Sander, O., Sommer, I., and Lengauer, T., Local protein structure prediction using discriminative models, *BMC Bioinformatics*, 7:14, 2006.
[26] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer-Verlag, 2000.
[27] Vlasov, P.K., Vlasova, A.V., Tumanyan, V.G., and Esipova, N.G., A tetrapeptide-based method for polyproline II-type secondary structure prediction, *Proteins*, 61:763–768, 2005.
[28] Wang, Y., Xue, Z.D., and Xu, J., Better prediction of the location of alpha-turns in proteins with support vector machine, *Proteins*, 65:49–54, 2006.
[29] Wang, Y., Xue, Z.D., Shi, X.H., and Xu, J., Prediction of pi-turns in proteins using PSI-BLAST profiles and secondary structure information, *Biochem. Biophys. Res. Commun.*, 347(3):574–580, 2006.
[30] Zhong, W., He, J., Harrison, R., Thai, P.C., and Pan, Y., Clustering support vector machines for protein local structure prediction, *Expert Syst. Appl.*, 32:518–526, 2007.
[31] `http://www.predictioncenter.org/casp8/`