Journal of the Korean Physical Society, Vol. 45, No. 6, December 2004, pp. 1441∼1449

Prediction of the Secondary Structures of Proteins by Using PREDICT, a Nearest Neighbor Method on Pattern Space

Keehyoung Joo, Ilsoo Kim, Seung-Yeon Kim and Jooyoung Lee*
School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-650

Julian Lee
Department of Bioinformatics and Life Science, Soongsil University, Seoul and Bioinformatics and Molecular Design Technology Innovation Center, and Computer Aided Molecular Design Research Center, Soongsil University, Seoul 156-743

Sung Jong Lee
Department of Physics and Center for Smart Bio-Materials, The University of Suwon, Suwon 445-890

(Received 10 September 2004)

We introduce a novel method for predicting the secondary structure of proteins, PREDICT (PRofile Enumeration DICTionary), in which the nearest-neighbor method is applied to a pattern space. For a given protein sequence, PSI-BLAST is used to generate a profile that defines patterns for amino acid residues and their local sequence environments. By applying the PSI-BLAST to protein sequences with known secondary structures, we construct pattern databases. The secondary structure of a query residue of a protein with unknown structure can be determined by comparing the query pattern with those in the pattern databases and selecting the patterns close to the query pattern. We have tested the PREDICT on the CB513 set (a set of 513 non-homologous proteins) in three different ways. The first test was based on a pattern database derived from 7777 proteins in the Protein Data Bank (PDB), including those homologous to proteins in the CB513 set and gave an average Q_3 score of 78.8 % per chain. In the second test, in order to carry out a more stringent benchmark test on the CB513 set, we removed from the 7777 proteins all proteins homologous to the CB513 set, leaving 4330 proteins. Pattern databases were constructed based on these proteins, and the average Q_3 score was 74.6 %. In the third test, we selected one query protein among the CB513 set and built pattern databases by using the remaining 512 proteins. This procedure was repeated for each of the 513 proteins, and the average Q_3 score was 73.1 %. Finally, we participated in the CASP5 (group ID: 531) where we employed the first-layer database based on the 7777 proteins and the second-layer database based on the CB513 set. The PREDICT gave quite promising results with an average Q_3 (Sov) score of 78.1 (77.4) % on 55 CASP5 targets.

PACS numbers: 05.10.−a, 42.30.Sy, 89.75.Kd, 87.14.Ee
Keywords: Protein structure prediction, Secondary structure prediction

I. INTRODUCTION

Determining the tertiary structure of a protein is very important in understanding the function and the biological role of the protein. The exponential growth in protein-sequence databases during recent years has by far outpace the experimental determination of the tertiary structures. Therefore, in the field of protein-structure investigation, it has become increasingly more popular to resort to computational methods [1–6] as an approach complementary to experimental structure determination. However, *ab-initio* predictions of the tertiary structures based solely on sequence information have not been successful so far [1, 2]. For this reason, many research efforts have been to devoted to determining the protein's secondary structure [2]. Reliable prediction of the secondary structure of a protein can serve as an intermediate step toward determining its tertiary structure [1, 2, 6].

Methods for secondary structure prediction have evolved from the early methods of single residue statistics [7], gradually incorporating correlation properties between neighboring residues [8]. In recent years, most of the successful methods have been based on the exploitation of evolutionary information [9–12] via sophisticated sequence alignment methods that can probe (oth-

*E-mail: jlee@kias.re.kr

erwise not-easily detectable) distant relationships between proteins. The most notable example is the use of the PSI-BLAST [13] to generate the profile, Position Specific Scoring Matrix, which is a numerical representation of the position-dependent substitution of amino acid residues. The profile elements inside a window of a fixed size centered on a given residue position can be used to define a pattern that describes the local sequence environment of that particular residue [11]. Then, the secondary structure of the residue can be predicted by investigating the pattern. One method to perform this task is to employ pattern recognition methods, such as neural networks [14,15], to predict the secondary structure elements of a given protein sequence [9,11,12].

Another line of approach to secondary structure prediction is the nearest-neighbor methods pursued by several groups [10,16,17], who utilized a distance measure based on the similarity of local configurations (nearby amino-acid-sequence identities) of residues. For example, Yi and Lander [16] proposed a nearest-neighbor method based on a scoring system that combined a sequence similarity matrix with a local structural environmental scoring method. In spite of these efforts, we believe that the potential of the nearest-neighbor approach has not been exploited to its full capacity. For example, sophisticated profile-generating algorithms, such as the PSI-BLAST, were not used in the early works.

In this work, we propose a method for predicting the secondary structures of proteins, PREDICT (PRofile Enumeration DICTionary), where we apply the nearest-neighbor approach to the patterns generated by PSI-BLAST. To the best of our knowledge, this is the first time for the nearest-neighbor idea to be combined with the PSI-BLAST to give a powerful algorithm for generating profiles. From these profiles, we can define patterns that represent the local sequence environment of residues. PREDICT directly probes the geometrical structure of the pattern space as represented by a collection of patterns, which we call a pattern database, corresponding to all the residues of proteins with known secondary structures. We begin with a distance measure defined in the pattern space. The distance measure between two patterns is defined as the weighted sum of the absolute values of the differences between the corresponding elements of the two patterns. The underlying idea of PREDICT is that patterns located close to each other in pattern space should have an identical secondary structure if we use a sensible definition of the distance measure.

Our method also has several new ingredients. First, we construct our own pattern database of 7777 proteins especially selected from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB), to be used for the secondary structure prediction. Second, we introduce the concept of a second-layer calculation to the nearest-neighbor approach, which has been used mainly in the context of neural network methods [11]. Finally, we optimize the weight parameters defining the distance measure in pattern space by using a set of proteins with known structures.

We tested PREDICT on the CB513 set [12] (a set of 513 non-homologous proteins; http://www.compbio.dundee.ac.uk/∼www-jpred/data/) and *participated* in the CASP5 [2] (http://predictioncenter.llnl.gov/casp5/pubResultS) (group ID: 531) for blind tests. The results of PREDICT are quite promising.

II. METHODS

1. The First-layer Calculation

Considering only the first chains, we obtained a set of 7777 proteins from the PDB after removing identical protein sequences. The structures of these proteins had all been determined by using X-ray crystallography with a resolution better than 3.0 Å. The secondary structures of these proteins are determined using the DSSP (Dictionary of Protein Secondary Structure) [18] routine. Following the CASP standard, we reduced the original eight-state classification into three-state one (G, H, I → H: helix; B, E → E: extended; T, S, _ → C: coil).

The profiles of these proteins were generated by using a PSI-BLAST search (version 2.2.4, with default option $E = 0.001$ and three iterations) for each target protein against the National Center for Biotechnology Information nonredundant sequence database (ftp://ncbi.nlm.nih.gov/blast/db/). Each profile consisted of a matrix of size $N_{\rm seq} \times 20$, where $N_{\rm seq}$ was the length of the protein sequence and 20 corresponded to the 20 amino-acid types. The elements of the profiles came from the relative frequencies of the amino acids for each residue position observed in the multiple sequence alignment.

We then defined the pattern for each residue by considering seven neighboring residues to the left and to the right of a given residue position, so that the size of the window was 15. Thus, a pattern constituted a 15×21 matrix where 21 stands for the 20 amino-acid types plus one indicating the vacancies at the N- and C-terminal ends of the protein sequences. The database of patterns generated from the set of 7777 proteins contains 1988085 patterns corresponding to all the residues in this set. We call this database the first-layer database.

For comparison, we also constructed first-layer databases based on reduced sets of proteins. One of them was derived from the set of 4330 proteins obtained by removing from 7777 proteins those homologous to the members of the CB513 set with a SD score [12] higher than 5. The other one was constructed by selecting 512 proteins from the CB513 set, which were then used for the test prediction of one remaining protein in the CB513 set (See Results and Discussion).

The distance between two patterns is defined by

$$D_{ij} = \sum_k W_k |P_i^{(k)} - P_j^{(k)}|, \quad (1)$$

where $P_i^{(k)}$ $(k = 1, 2, \ldots, 15 \times 21)$ is the k-th component of the pattern i, and $\{W_k\}$ are the weight parameters. Since we expect the pattern components nearer to the center residue to be more important in defining the distance, we use as an initial guess for the weights $W_k = (8 - |8 - r|)^2$, where r $(r = 1, 2, \ldots, 15)$ is the index labeling the residue position corresponding to the k-th component. (For example, $r = 8$ corresponds to the center residue.) We denote this parameter set as $\{W_k^0\}$.

Since we expect that the patterns close to one another in the pattern space to share the identical secondary structure, we enumerate all pairwise distances between a query pattern and the patterns in the database, and select the N-nearest patterns. We then simply count the occurrences of helix ('H'), extended ('E'), and coil ('C') patterns among these N patterns. The secondary structure of a query pattern can be simply determined by the majority rule in which the secondary structure of the most occurrences is chosen for the prediction. We call this procedure as a first-layer calculation. The cutoff number N can be chosen suitably by trial and error. We have used various values of N (See Results and Discussion).

2. The Second-layer Calculation

A first-layer calculation alone may be used for predicting the secondary structure of proteins. But we found that we can significantly improve the efficiency of the prediction by introducing the so-called second-layer calculation, which is analogous to the one used in second-level neural networks [11]. Namely, instead of applying the majority rule from the occurrences of 'H', 'E', and 'C' one can construct another kind of pattern based on the first-layer calculation in the following way:

For a given sequence, its profile is generated by using PSI-BLAST, and the pattern for each residue is obtained following the procedure described earlier. Then, the first-layer enumeration is performed by comparing the query pattern with those in the first-layer pattern database, as described in the previous section.

The result of the first-layer calculation provides us a three-state (H, E, C) frequency table for each query residue. These frequency tables provide us with another kind of pattern by considering a window of size 15 on each residue. We call the resulting pattern the second-layer pattern, which consists of 15×4 elements, where the additional fourth column denotes vacancies at the terminal ends of the sequences. Therefore, by performing the first-layer calculations for the protein residues whose secondary structures are known, we construct the database of the second-layer patterns, which we call the second-layer pattern database.

In order to perform the second-layer calculation, we first perform the first-layer calculation for a query residue to obtain the corresponding second-layer pattern. Then the calculation is performed in a fashion similar to that used for the first-layer one by comparing the query second-layer pattern with those of the second-layer pattern database. The distance measure in the second-layer pattern space is defined in a similar way as in the case of the first-layer pattern space with a trivial difference in the number of columns (4 vs. 21):

$$D_{ij} = \sum_k \tilde{W}_k |S_i^{(k)} - S_j^{(k)}|, \quad (2)$$

where $S_i^{(k)}$ $(k = 1, 2, \ldots, 15 \times 4)$ is the k-th component of the pattern i, and $\{\tilde{W}_k\}$ are the weight parameters. We again use weights of $\tilde{W}_k = (8 - |8 - r|)^2$, where r $(r = 1, 2, \ldots, 15)$ is the index labeling the position of the residue corresponding to the k-th component. Using this distance matrix, we select the N-nearest patterns for the query residue. Then, the secondary structure of the query residue is predicted following the majority rule and using these N-nearest patterns. We used the same values of N that were used in the first-layer calculation. We find that applying the second-layer procedure improves the performance of the prediction substantially over the one based on the first-layer method alone (See Results and Discussion).

3. Distance Measure and Weight Optimization

For the distance between two first-layer patterns, Eq. (1), we use an initial guess for the the weights $W_k^0 = (8 - |8 - r|)^2$, where r $(r = 1, 2, \ldots, 15)$ is the index labeling the residue position corresponding to the k-th component. Obviously, the parameter set $\{W_k^0\}$ is not the best one, and it would be desirable to seek a better weight parameter set. Therefore, we optimize the parameters so that the success rate of our prediction increases for a given set of proteins, which is the CB513 set in this work. We first perform the first-layer calculation for each residue in this set by using all the other residues in the CB513 set to generate a pattern database. The method of prediction used here for evaluating the performance of the parameter set is slightly different from the one described in other sections. We select three sets of 100 nearest patterns whose secondary structures are H, E, and C. We calculate the average distances between the patterns in each set and the query pattern, denoted by D_H, D_E, and D_C. We will use the secondary structure corresponding to the least value among D_H, D_E, and D_C as the prediction. The fraction of residues of the proteins in the CB513 set whose secondary structure is correctly predicted using the initial parameter set $\{W_k^0\}$ is 71.0 %.

In order to optimize the parameters, we first define the gaps g_1 and g_2 as follows. Suppose that the secondary structure of a given query residue is a helix. In this case, the gaps are defined as

$$g_1 = D_H - D_C,$$
$$g_2 = D_H - D_E \ .$$

For residues with secondary structures of E or C, the gaps are defined in a similar way. It is obvious that for a residue whose experimental secondary structure agrees with the prediction, both g_1 and g_2 are negative. Here, we want to change the parameters so that many residues in the CB513 set are predicted correctly. We first select residues whose correct secondary structures differ from the prediction, and call the resulting subset as set A. We define $g = \max(g_1, g_2)$ and choose the lower 10 % (in the value of g) of the residues in the set A. We call the resulting subset as set B. The residues in set B are considered as the ones whose gaps can be easily converted into negative values by parameter optimization. We then minimize g for all residues in set B, one by one. For each residue, the gaps are linear functions of the weight parameters,

$$g_1 = \sum_k W_k d_1^k \ ,$$

and

$$g_2 = \sum_k W_k d_2^k \ ,$$

where the components d_j^k can be easily calculated from the pattern elements. If $g_1 > g_2$, we increase the parameters $\{W_k\}$ by an amount δW_k:

$$\delta W_k = -\epsilon \ \mathrm{sign}(d_1^k) W_k \ ,$$

where ϵ is a small positive number. Similarly, for $g_1 < g_2$,

$$\delta W_k = -\epsilon \ \mathrm{sign}(d_2^k) W_k \ .$$

We repeat this procedure 50 times for each residue in set B. When all the residues in set B have been used for parameter optimization, one iteration is completed. We start the next iteration by evaluating the gaps of all the residues in the CB513 set, selecting the residues with incorrect secondary structure calculation results, selecting the 10 % among them with the smallest gaps, and minimizing these gaps. We perform 300 iterations and call the resulting parameter set as $\{W_k^{300}\}$. We use $\epsilon = 0.2/N_{patt}$, where $N_{patt} = 84119$ is the number of patterns in the CB513 set. After the parameter optimization, the fraction of residues with negative gaps is indeed, increased to $Q_3 = 73.1$ % from the initial value of $Q_3 = 71.0$ % (See Results and Discussion for the definition of Q_3).

It should be noted that as we change the parameters to minimize gaps for a particular residue, the gaps for other residues might increase as a result. For this reason, we use a very small value of ϵ, which is inversely proportional to N_{patt}. Of course, to treat the problem more rigorously, we might consider optimizing gaps for a particular residue while imposing (linear) constraints on the gaps of other residues. This results in an optimization problem called linear programming, where the object function and the constraints are all linear functions. We could not pursue this line of investigation due to the large computer memory requirement. Also, in principle, one might consider a set of proteins that is more extensive than the CB513 set, which would require more extensive computational resources.

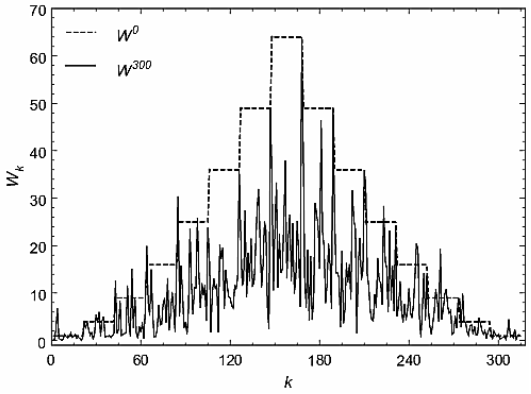

Fig. 1. Values of the parameters W_k for $k = 1, 2, \ldots, 15 \times 21$ are plotted for $\{W_k^0\}$ and $\{W_k^{300}\}$. We note that the parameter set $\{W_k^{300}\}$ depends on the amino-acid type and on the position from the center residue.

Since the parameter optimization was performed using only the first-layer calculation on the CB513 set, $\{W_k^{300}\}$ is used only in the steps of the first-layer enumeration. We show the values of the parameter sets $\{W_k^0\}$ and $\{W_k^{300}\}$ in Fig. 1. We note that in contrast to the original weights $\{W_k^0\}$, which depend only on the position of the residue corresponding to the pattern element k from the central residue, the optimized weights exhibit an explicit additional dependence on the amino-acid type.

III. RESULTS AND DISCUSSION

In this section, we use the Q_3 score, Sov (the Segment OVerlap) measure [19], and the Matthews correlation coefficients [20] to evaluate the performance of PREDICT. Q_3 gives percentage of residues predicted correctly for all three conformational states (H,E,C):

$$Q_3 = \frac{N_{\mathrm{cor}}}{N_{\mathrm{res}}}, \tag{3}$$

where N_{res} is the total number of residues, and N_{cor} is the number of residues whose conformational states are predicted correctly. Q_3 in Eq. 3 can be calculated for all the residues of the target proteins, denoted as Q_3^r, or alternatively, it can be calculated for each protein se-

quence and be averaged over the set of protein chains with equal weights, which we denote as Q_3^c.

In order to define the Sov measure, we define the set of overlapping segments with secondary structure i (= H,E,C):

$$S(i) = \{(s_1(i), s_2(i)) | s_1(i) \cap s_2(i) \neq \emptyset\}, \quad (4)$$

where ($s_1(i)$ and $s_2(i)$) are pairs of observed and predicted secondary structure segments in conformational state i (= H,E,C), which has at least one residue in common. $S(i)$ is the set of all such pairs. Then, the Sov measure is defined as

$$\mathrm{Sov} = \frac{1}{N_{\mathrm{pair}}} \sum_{i \in \{H,E,C\}} \sum_{S(i)} \left[\frac{\mathrm{minov}(s_1(i), s_2(i)) + \delta(s_1(i), s_2(i))}{\mathrm{maxov}(s_1(i), s_2(i))} \cdot \mathrm{len}(s_1(i)) \right], \quad (5)$$

where $\mathrm{len}(s_1(i))$ and $\mathrm{len}(s_2(i))$ are the numbers of residues in the segments $s_1(i)$ and $s_2(i)$, respectively, $\mathrm{minov}(s_1(i)\mathrm{and}s_2(i))$ are the lengths of actual overlap of a given pair $s_1(i)$ and $s_2(i)$, $\mathrm{maxov}(s_1(i)\mathrm{and}s_2(i))$ are the lengths of the total extent of the residues which belong to either $s_1(i)$ or $s_2(i)$, and

$$\delta(s_1, s_2) = \min[(\mathrm{maxov}(s_1, s_2) - \mathrm{minov}(s_1, s_2)) ; \mathrm{minov}(s_1, s_2); \mathrm{int}(\mathrm{len}(s_1)/2); \mathrm{int}(\mathrm{len}(s_2)/2)] \quad (6)$$

Also, the normalization factor N_{pair} is defined as:

$$N_{\mathrm{pair}} = \sum_{i \in H,E,C} \left[\sum_{S(i)} \mathrm{len}(s_1(i)) + \sum_{S'(i)} \mathrm{len}(s_1(i)) \right], \quad (7)$$

where $S'(i)$ is the set of observed segments $s_1(i)$ that has no overlap with the predicted segments of the secondary structure i.

The Matthews correlation coefficients are defined as follows: For each of the conformational state H, E, and C,

$$C_i = \frac{(p_i n_i) - (u_i o_i)}{\sqrt{(n_i + u_i)(n_i + o_i)(p_i + u_i)(p_i + o_i)}}, \quad (8)$$

where p_i is the number of correctly predicted residues in conformational state i (= H,E,C), n_i is the number of residues that are correctly identified as something other than state i, o_i is the number of residues that are not in state i but are incorrectly predicted as in state i, and u_i is the number of residues in state i that are missed by the algorithm.

1. CB513 Set

In order to test the performance of PREDICT, we chose the 513 non-homologous proteins of the CB513 set. We carried out three benchmark tests on the CB513 set. In the first benchmark test (benchmark I), we used all 7777 sequences to construct the first-layer database and used the 513 non-homologous protein sequences of the CB513 set to construct the second-layer database.

Since the 7777 sequences include those which are homologous to the sequences in the CB513 set, we removed them from the 7777 sequences. For this calculation, we used the SD score defined by Cuff and Barton [12] as the criterion for homology. This score is known to be more rigorous than the usual criterion based on sequence identity [12]. In the 7777 sequences, those with SD scores above 5.0 against any sequence in the CB513 set were eliminated to obtain a set of 4330 sequences. In the second benchmark test (benchmark II), we used these 4330 sequences to construct the first- and the second-layer databases.

Finally, in the third benchmark test (benchmark III), we constructed the first- and the second-layer databases consisting of 512 sequences by excluding a test sequence from the CB513 set, and we carried out the benchmark prediction for this test sequence. This procedure was repeated for each of the 513 sequences in the CB513 set. It should be noted that for each of the 513 calculations, we constructed separate databases of the first and the second layers.

It should be noted that, in contrast to benchmark tests II and III, the database for benchmark test I contains sequences homologous to the query protein. In fact, in an actual blind prediction such as CASP, we intend to use any possible homology between the query sequence and those in the database, so benchmark I estimates the performance of PREDICT in such a situation, where the performance is expected to be improved by using all 7777 proteins in the database.

In benchmark I, we paid special attention to exclude the patterns coming from the query protein sequence in the pattern database. The sequences in the CB513 set were also used for optimizing parameters defining the distance measure in pattern space. (One should not confuse this procedure with the training procedure in neural network-based method) The parameters optimized using the CB513 set are called $\{W_k^{300}\}$ whereas the unoptimized parameters are called $\{W_k^0\}$ (See Methods). The

Table 1. The PREDICT predictions of benchmark I for the CB513 set. Q_3^r and Q_3^c are the Q_3 scores obtained by averaging over 84199 residues and 513 chains, respectively. Also, C_H, C_E, and C_C denote Matthews correlation coefficients for H, E, and C, respectively, The column labelled as W^0 shows the results obtained using the initial parameter set, and the one below the label W^{300} shows the results obtained using the optimized parameters in the first-layer calculation (see text). The number N denotes the number of closest patterns utilized for the prediction by the majority rule. We use the first-layer database based on the 7777 proteins and the second-layer database based on the CB513 set.

first layer												
	W^0						W^{300}					
N	Q_3^r	Q_3^c	Sov	C_H	C_E	C_C	Q_3^r	Q_3^c	Sov	C_H	C_E	C_C
1	88.1	84.7	79.0	0.75	0.74	0.72	88.3	85.1	78.8	0.76	0.75	0.72
10	82.1	79.4	71.5	0.65	0.63	0.63	82.6	80.1	71.0	0.67	0.63	0.64
20	79.0	76.9	68.0	0.60	0.58	0.58	79.7	77.7	68.3	0.62	0.59	0.60
30	77.8	75.9	67.8	0.58	0.56	0.56	78.3	76.5	67.1	0.60	0.56	0.57
40	76.7	75.0	67.0	0.57	0.55	0.55	77.5	75.9	66.8	0.59	0.55	0.56
50	76.0	74.5	66.3	0.56	0.53	0.54	76.9	75.5	66.3	0.58	0.54	0.56
60	75.4	74.1	66.1	0.55	0.53	0.54	76.4	75.0	66.2	0.57	0.54	0.55
70	75.1	73.8	65.9	0.54	0.52	0.53	76.1	74.7	66.0	0.57	0.53	0.54
80	74.8	73.6	65.8	0.54	0.51	0.53	75.9	74.7	66.0	0.57	0.52	0.54
90	74.5	73.4	65.5	0.53	0.51	0.52	75.7	74.5	66.2	0.57	0.52	0.54
100	74.3	73.3	65.9	0.53	0.51	0.52	75.4	74.3	66.3	0.56	0.52	0.54
200	73.1	72.2	65.5	0.51	0.49	0.50	74.3	73.4	65.8	0.55	0.50	0.52
300	72.3	71.4	65.1	0.50	0.48	0.49	73.6	72.8	65.7	0.54	0.49	0.51
400	71.8	70.9	64.8	0.49	0.47	0.48	73.2	72.3	65.7	0.53	0.48	0.50
500	71.6	70.8	64.7	0.49	0.48	0.48	73.0	72.0	65.5	0.52	0.48	0.50

second layer												
	W^0						W^{300}					
N	Q_3^r	Q_3^c	Sov	C_H	C_E	C_C	Q_3^r	Q_3^c	Sov	C_H	C_E	C_C
1	91.8	90.3	86.1	0.84	0.83	0.81	90.2	87.8	83.0	0.81	0.78	0.76
10	89.8	87.3	84.1	0.80	0.76	0.76	89.9	87.7	84.3	0.81	0.76	0.76
20	86.9	84.6	81.6	0.75	0.70	0.71	87.2	85.1	82.1	0.76	0.70	0.71
30	85.5	83.4	80.4	0.73	0.68	0.68	85.8	83.8	80.9	0.74	0.68	0.69
40	84.6	82.6	79.7	0.72	0.67	0.67	85.0	83.0	80.1	0.72	0.67	0.68
50	83.9	81.9	79.1	0.70	0.66	0.65	84.3	82.5	79.3	0.71	0.66	0.67
60	83.4	81.4	78.5	0.69	0.65	0.65	83.6	81.9	78.7	0.70	0.65	0.66
70	82.8	80.8	77.8	0.68	0.64	0.64	83.0	81.3	78.3	0.69	0.64	0.64
80	82.3	80.5	77.5	0.67	0.63	0.63	82.5	80.9	78.0	0.69	0.63	0.64
90	81.6	79.8	77.0	0.66	0.62	0.62	81.9	80.4	77.4	0.68	0.62	0.63
100	80.3	78.8	75.9	0.64	0.60	0.60	80.3	78.8	75.9	0.64	0.60	0.60
200	77.9	76.8	74.0	0.61	0.56	0.56	77.9	76.8	74.0	0.61	0.56	0.56
300	76.5	75.5	72.9	0.58	0.54	0.54	76.5	75.5	72.9	0.58	0.54	0.54
400	75.8	74.8	72.3	0.57	0.52	0.53	75.8	74.8	72.3	0.57	0.52	0.53
500	75.2	74.4	71.9	0.57	0.51	0.52	75.2	74.4	71.9	0.57	0.51	0.52

benchmark results for the CB513 set are shown in Table I. In the tables, N denotes the number of closest patterns utilized for the prediction by using the majority rule, and when $\{W_k^{300}\}$ is used, it is used only for the first-layer calculation (See Methods). All second-layer calculations were carried out using $\{W_k^0\}$. Among the Q_3 scores, Q_3^r is the one obtained by averaging over the total number of residues (84199) while Q_3^c is the one obtained by averaging over the number of chains (513). The slightly larger values of Q_3^r over those of Q_3^c are due to the fact that results for smaller-sized (less than 100 amino acids)

Table 2. Results of the benchmark test II. The first-layer and the second-layer databases are constructed based on the 4330 proteins which are not homologous to the CB513 set.

benchmark II												
layer	first layer						second layer					
N	Q_3^r	Q_3^c	Sov	C_H	C_E	C_C	Q_3^r	Q_3^c	Sov	C_H	C_E	C_C
1	69.3	68.8	57.9	0.48	0.45	0.44	69.5	69.1	58.5	0.48	0.45	0.42
10	70.4	70.0	59.6	0.49	0.45	0.46	73.6	73.3	67.9	0.55	0.51	0.50
20	70.8	70.4	61.0	0.49	0.45	0.47	74.3	73.9	69.2	0.56	0.52	0.51
30	71.0	70.4	61.7	0.49	0.46	0.47	74.4	74.1	70.1	0.56	0.52	0.52
40	71.0	70.3	61.9	0.49	0.46	0.47	74.5	74.1	70.7	0.56	0.52	0.52
50	71.0	70.4	62.2	0.49	0.46	0.47	74.5	74.1	70.6	0.56	0.52	0.52
60	71.0	70.3	62.1	0.49	0.45	0.47	74.5	74.1	70.7	0.56	0.52	0.52
70	71.2	70.5	62.6	0.49	0.45	0.48	74.6	74.1	70.7	0.56	0.52	0.52
80	71.3	70.6	62.9	0.49	0.45	0.48	74.6	74.2	70.8	0.56	0.52	0.52
90	71.3	70.6	63.2	0.49	0.45	0.48	74.6	74.1	70.9	0.56	0.52	0.52
100	71.6	70.8	63.8	0.50	0.46	0.48	74.6	74.2	71.0	0.56	0.52	0.52
200	71.7	70.9	64.9	0.50	0.46	0.48	74.7	74.2	71.3	0.57	0.52	0.52
300	71.7	70.9	65.3	0.50	0.47	0.48	74.7	74.3	71.5	0.57	0.52	0.52
400	71.5	70.7	65.3	0.49	0.47	0.48	74.7	74.2	71.4	0.57	0.51	0.52
500	71.4	70.7	65.7	0.49	0.47	0.48	74.7	74.2	71.4	0.57	0.51	0.52

Table 3. Results of the benchmark test III. In this test, one protein in the CB513 is used in turn as a query protein, and the first-layer and the second-layer databases are constructed based on the remaining 512 proteins of the CB 513 set. This procedure was repeated 513 times, once for each protein in the set.

benchmark III												
layer	first layer						second layer					
N	Q_3^r	Q_3^c	Sov	C_H	C_E	C_C	Q_3^r	Q_3^c	Sov	C_H	C_E	C_C
1	61.8	61.6	49.4	0.35	0.33	0.31	65.2	64.6	53.1	0.40	0.37	0.34
10	69.1	68.8	59.7	0.46	0.44	0.44	72.0	71.6	65.1	0.52	0.47	0.47
20	70.1	69.8	62.6	0.47	0.45	0.46	72.9	72.5	67.7	0.53	0.49	0.49
30	70.5	70.1	63.7	0.48	0.45	0.46	73.2	72.8	68.6	0.54	0.49	0.49
40	70.6	70.2	64.5	0.48	0.46	0.46	73.4	72.8	68.8	0.54	0.49	0.50
50	70.7	70.3	64.7	0.48	0.46	0.47	73.4	72.9	69.1	0.54	0.49	0.50
60	70.7	70.3	64.9	0.48	0.46	0.47	73.5	73.0	69.4	0.55	0.49	0.50
70	70.8	70.3	65.1	0.48	0.46	0.47	73.5	73.0	69.5	0.54	0.49	0.50
80	70.8	70.3	65.3	0.48	0.46	0.47	73.5	73.0	69.4	0.54	0.49	0.50
90	70.9	70.4	65.3	0.48	0.46	0.47	73.5	73.0	69.6	0.55	0.49	0.50
100	70.9	70.4	65.4	0.48	0.46	0.47	73.5	73.1	69.7	0.55	0.49	0.50
200	70.8	70.4	65.8	0.48	0.46	0.47	73.4	72.9	69.7	0.54	0.49	0.50
300	70.6	70.0	65.5	0.47	0.45	0.47	73.4	72.8	69.6	0.55	0.49	0.50
400	70.5	69.9	65.0	0.47	0.45	0.46	73.3	72.7	69.5	0.55	0.49	0.49
500	70.3	69.7	65.1	0.46	0.45	0.46	73.2	72.6	69.4	0.54	0.49	0.49

sequences are poorer than those for bigger-sized ones.

We observe that the performance of the second-layer prediction is consistently better than that of the first-layer one by about 3 – 10 %. It should be noted that, even with the unoptimized parameters $\{W_k^0\}$, the performance of PREDICT is quite excellent. We also apply the optimized parameter set $\{W_k^{300}\}$ in order to check whether the parameter was optimized properly, and we find the efficiency of the parameter set $\{W_k^{300}\}$ for defining the distance measure to be only slightly better than that of $\{W_k^0\}$. The histograms for the Q_3 and the Sov scores for $N = 100$ are shown in Fig.2. We use the op-

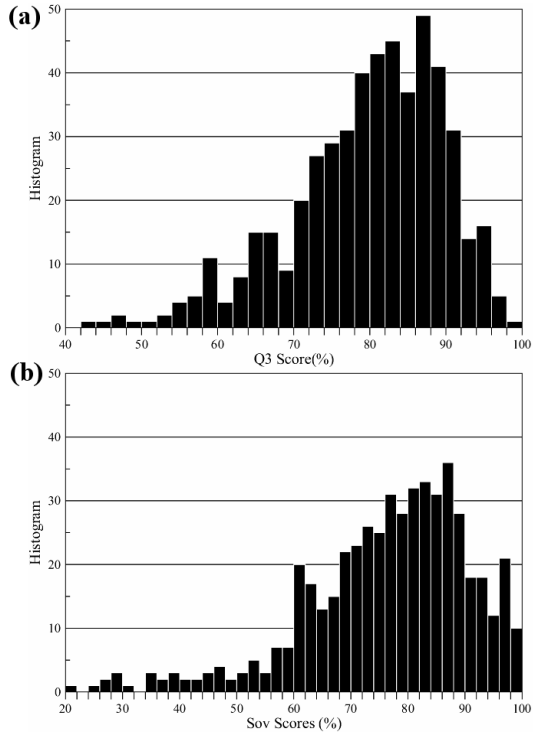

Fig. 2. Second-layer prediction (with parameters $\{W_k^{300}\}$ in the first-layer calculation) for the CB513 set: (a) histogram of Q_3 scores for $N = 100$, and (b) histogram of Sov scores for $N = 100$.

timized parameter set $\{W_k^{300}\}$ for the predictions of the CASP5 targets (see the next subsection).

We tried various values of N and found that using $N = 1$ gave the best performance for benchmark I. However, we believe this result is due to the fact that there are proteins in the database which are homologous to the query protein; thus using $N = 1$ corresponds to selecting the most homologous protein. In fact, in the case of benchmark tests II and III (see below), where proteins homologous to the query protein were removed from the database, the performance of PREDICT did not improve as we decreased the value of N below around 100.

For a more stringent evaluation of our calculations, we used exclusively $\{W_k^0\}$ for all benchmark tests below (benchmarks II and III of Tables II and III). In benchmark II, using the set of 4330 sequences, the best performance for the second-layer calculation with $N = 200$, where $Q_3^r = 74.7$, $Q_3^c = 74.3$, Sov = 71.5, $C_H = 0.57$, $C_E = 0.52$, and $C_C = 0.52$. The CPU time for the prediction was less than 10 seconds per residue on average, using an AMD MP2200 1.8 GHz CPU. In benchmark III, where the databases were constructed separately for each sequence in the CB513 set, the best performance was obtained for the second-layer calculation with $N = 100$, where $Q_3^r = 73.5$, $Q_3^c = 73.1$, Sov = 69.7, $C_H = 0.55$, $C_E = 0.49$, and $C_C = 0.50$.

2. CASP5 Targets

In order to obtain the performance of PREDICT in a blind test on sequences with unknown structures, we *participated* in CASP5 (http://predictioncenter.llnl.gov/casp5/pubResultS) (group ID:531) and applied the the PREDICT procedure to the CASP5 targets. We used the optimized parameter set $\{W_k^{300}\}$ with $N = 100$, and we used the CB513 set for the second-layer pattern database due to the time constraint for submiting our results to CASP5 on time. The average Q_3^r score was 77.55 %, the average Q_3^c score for 55 targets was 78.09 % with standard deviation of 7.24 %, and the average Sov score was 77.38 % with the standard deviation of 9.75 %; the average Matthews coefficients for H, E, and C were 0.68, 0.65, and 0.58, with standard deviations 0.14, 0.17, and 0.12, respectively. After CASP5, we repeated the calculation using the 7777 proteins, instead of CB513 set, as the second-layer pattern database, and we obtained an average Q_3^r score of 77.66 %, an average Q_3^c score of 78.12 % with standard deviation of 7.66 %, and an average Sov score of 77.63 % with a standard deviation of 10.30 %; the average Matthews coefficients for H, E, and C were 0.68, 0.65, and 0.59 with standard deviations of 0.14, 0.17, and 0.13, respectively. We observe that the performance of PREDICT improves only slightly as we expand the size of the second-layer pattern database.

IV. CONCLUSION

We have presented a new method for predicting the secondary structure of proteins, PREDICT, based on the concept of the distance measure in the pattern space. To the best of our knowledge, this is the first time that the nearest-neighbor idea has been applied to the patterns generated from PSI-BLAST. The results from the CB513 set and the 55 CASP5 targets have shown that the performance of PREDICT is quite promising.

It would also be interesting to probe the geometry of the entire pattern space with respect to the distance measure and to the many islands of the three subspaces corresponding to the 'H', 'E', and 'C'. This would require an understanding of the local and the global-clustering properties of these subspaces. In addition, the effective dimensions for these subspaces should be investigated. Understanding these features should lead to further improvements in PREDICT.

ACKNOWLEDGMENTS

This work was supported by grant No. R01-2003-000-11595-0 (Sung Jong Lee and Jooyoung Lee) and No. R01-2003-000-10199-0 (Julian Lee) from the Basic

Research Program of the Korea Science & Engineering Foundation.

REFERENCES

[1] D. Baker and A. Sali, Science **294**, 93 (2001).
[2] P. Aloy, A. Stark, C. Hadley and B. R. Russel, Proteins **53**, 436 (2003).
[3] K. Joo, J. Lee, S.-Y. Kim, I. Kim, J. Lee and S. J. Lee, J. Korean Phys. Soc. **44**, 599 (2004).
[4] J. Sim, S.-Y. Kim, J. Lee and A. Yoo, J. Korean Phys. Soc. **44**, 611 (2004).
[5] M. Heo, S. Kim, E.-J. Moon, M. Cheon, K. Chung and I. Chang, J. Korean Phys. Soc. **44**, 1571 (2004).
[6] J. Lee, S.-Y. Kim, K. Joo, I. Kim and J. Lee, Proteins **56**, 704 (2004).
[7] P. Y. Chou and G. D. Fasman, Biochemistry **13**, 222 (1974).
[8] J. Garnier, D. J. Osguthorpe and B. Robinson, J. Mol. Biol. **120**, 97 (1978).
[9] B. Rost and C. Sander, J. Mol. Biol. **232**, 584 (1993).
[10] A. A. Salamov and V. V. Solovyev, J. Mol. Biol. **247**, 11 (1995).
[11] D. T. Jones, J. Mol. Biol. **292**, 195 (1999).
[12] J. A. Cuff and G. J. Barton, Proteins **34**, 508 (1999).
[13] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Nucl. Acids. Res. **25**, 3389 (1997).
[14] T. Shimizu, J. Korean Phys. Soc. **40**, 1072 (2002).
[15] S. Fujiki, M. Nakao and N. M. Fujiki, J. Korean Phys. Soc. **40**, 1091 (2002).
[16] T. M. Yi and E. S. Lander, J. Mol. Biol. **232**, 1117 (1993).
[17] J. M. Levin, Protein Eng. **10**, 771 (1997).
[18] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
[19] A. Zemla, C. Venclovas, K. Fidelis and B. Rost, Proteins **34**, 220 (1999).
[20] B. W. Matthews, Biochim. Biophys. Acta. **405**, 442 (1975).