PREDICTION REPORT

Predicting Protein Secondary Structure and Solvent Accessibility with an Improved Multiple Linear Regression Method

Sanbo Qin,[1,3] Yun He,[1,3] and Xian-Ming Pan[1,2]*

[1]*National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, People's Republic of China*

[2]*Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing, People's Republic of China*

[3]*Graduate School of the Chinese Academy of Sciences, Beijing, People's Republic of China*

ABSTRACT **We have improved the multiple linear regression (MLR) algorithm for protein secondary structure prediction by combining it with the evolutionary information provided by multiple sequence alignment of PSI-BLAST. On the CB513 dataset, the three states average overall per-residue accuracy, Q_3, reached 76.4%, while segment overlap accuracy, SOV99, reached 73.2%, using a rigorous jackknife procedure and the strictest reduction of eight states DSSP definition to three states. This represents an improvement of approximately 5% on overall per-residue accuracy compared with previous work. The relative solvent accessibility prediction also benefited from this combination of methods. The system achieved 77.7% average jackknifed accuracy for two states prediction based on a 25% relative solvent accessibility mode, with a Mathews' correlation coefficient of 0.548. The improved MLR secondary structure and relative solvent accessibility prediction server is available at http://spg.biosci.tsinghua.edu.cn/. Proteins 2005;61:473–480.**

Key words: protein secondary structure prediction; solvent accessibility prediction; multiple linear regression; protein folding; PSSM

INTRODUCTION

With the achievement of mapping the human genome and other large-scale genome sequencing projects,[1–5] efforts are underway to understand the roles of gene products (i.e., proteins) in biological pathways and human diseases and to exploit their functional roles. A key aspect in the "postgenomic" era will be the increasingly widespread use of protein structures. Although most of the genome is now available as sequence data, little is known about protein functions and, in turn, even less is known about individual protein structures. Furthermore, the sequence–structure gap is even more rapidly increasing. Currently, three-dimensional structures of less than one in 40 sequences have been determined, as estimated by comparing the number of entries in the PDB and Uniprot databases. Therefore, the prediction of protein tertiary structure and ultimately protein function from amino acid sequence is arguably one of the most important problems in molecular biology.[6]

Assessment from the CASP[7] and CAFASP[8] experiments demonstrated that the problem of protein structure prediction is still open. Some methods, such as homology modeling or threading are useful, and have produced advances in protein structure prediction, but are not always feasible. Because these methods rely heavily on structure/fold recognition, that is, identification of a protein with a known structure that shares the same fold with the target protein as a structural model (template), this has been done by identifying proteins that have similar sequences. However, there are many proteins that share the same fold but have no clear sequence similarity.

On the other hand, significant advances in the prediction of secondary structure and residue solvent accessibility have been seen over the last 40 years.[9] As the most familiar and well-defined problems, protein secondary structure and residue solvent accessibility are often regarded as the first step in protein tertiary structure prediction, and the prediction results can serve as inputs to generate a template for protein tertiary structure

Grant sponsor: Natural Science Foundation of China (NSFC); Grant numbers: 30230100 and 90103031; Grant sponsor: the 863 project; Grant number: 2004AA235100; Grant sponsor: the CNHLPP project (from the China Commission for Science and Technology); Grant number: 2004BA711A21

*Correspondence to: Xian-Ming Pan, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing, People's Republic of China. E-mail: xmpan@sun5.ibp.ac.cn

Received 31 January 2005; Accepted 16 May 2005

Published online 8 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20645

prediction. These methods are powerful tools in protein structure prediction from amino acid sequences.

Currently, most of the available successful protein secondary structure and solvent-accessibility prediction methods are based on neural network or machine-learning methods and take a nonlinear approach. They are "black-box" predictors; they do not make the basis of their predictions explicit, nor do they provide insight into the principles governing the formation of secondary structure. In contrast to this, in this article, we report an improved protein secondary structure and solvent-accessibility prediction method based on multiple linear regression.[10,11] By combining the position specific information provided by PSI-BLAST[12] and employing a second-layer prediction, on a test set of 513 protein chains (CB513), this method achieved an average per-residue overall accuracy score, Q_3, between 76.4 and 79.3%, depending on the precise definition of observed secondary structure used. This level of accuracy is similar to "state-of-the-art" methods, such as the PSIPRED[13] prediction method. By altering the structural states from secondary structure to relative solvent accessibility, this system was also used for solvent-accessibility prediction, and produced fairly good results.

MATERIALS AND METHODS

Dataset and Structure Definitions

A systematic testing of the performance is a precondition for any prediction program to be reliably useful. The use of a common data set is necessary when the prediction accuracy of different programs needs to be assessed in comparison to each other. In this study, the proteins used were a set of 513 chains (CB513) compiled by Cuff and Barton,[14] which is representative of high-resolution structures and was selected according to a stringent definition of sequence similarity.

The secondary structural states were defined by the DSSP program,[15] which provides an eight-state assignment of secondary structure. The complete dataset contains a total of 84,119 residues (including 28 unassigned residues treated as coil), with 30.9% H (alpha-helix), 3.7% G (3/10 helix), 21.3% E (extended beta-strand), 1.4% B (isolated strand), 11.9% T (turn), 9.8% S (bend), 21.1% coil, and only 0.03% I (pi helix). There are five different published eight- to three-state reduction methods.

1. Method A: H, G and I to H; E and B to E; Rest to C.[16]
2. Method B: H to H; E, B to E; Rest to C.[17]
3. Method C: H to H; E to E; Rest to C.[17]
4. Method D: GGGHHHH redefined as HHHHHHH, then B and GGG to C, with H to H and E to E, rest to C.[18]
5. Method E: H to H; E to E; Rest to C including EE and HHHH.[19]

In this article, reduction Method A was adopted unless otherwise explicitly stated, because it is considered to be the strictest definition, which usually results in lower prediction accuracy than others. This also allows comparison with most other prediction methods listed on the EVA server.[16]

Solvent-accessible surface information was extracted from the DSSP files of Kabsch and Sander.[15] Relative solvent accessibility was computed by normalizing the accessible surface area with the maximum values reported by Shrake and Rupley.[20] A 25% cutoff was used to define the two states of solvent accessibility (buried and exposed).

PSI-BLAST Profiles

The multiple sequence alignment profiles used in this study, in the form of a 20 × M (the target sequence length) position-specific scoring matrix, were generated by three iterations of PSI-BLAST, with an *E*-value cutoff at 0.001.[12] To scale these profiles to the required 0–1 range, the standard logistic function was used:

$$f(x) = \frac{1}{1 + \exp(-x)} \quad (1)$$

where x is the raw profile matrix value. This is different from the previous version of the MLR method, in which the frequency counts for each amino acid were used.

Algorithm

In previous studies,[10,11] we applied the multiple linear regression method (MLR) to protein secondary structure prediction and solvent-accessibility prediction. Here, we describe a dual-layer MLR system as follows:

We are interested in predicting the structural state of residue i, ω_i (secondary structure, solvent accessibility, or other), based on knowledge of the amino acid subsequence, $\{A_j\}$, within a "window" of restricted size n residues symmetric about location i. Locations within the window are indexed by j, with $I_X(\omega_i)$ defining the structural information. The value of $I_X(\omega_i)$ was taken as 1 when the residue in position i was in the state X; otherwise as 0. [For example, for solvent accessibility prediction, if the residue in position i is in a buried state (B), then $I_B(\omega_i) = 1$, and $I_E(\omega_i) = 0$; for secondary structure prediction, if the residue in position i is in a helix state (H), then $I_H(\omega_i) = 1, I_E(\omega_i) = 0$ and $I_C(\omega_i) = 0$]. The equation for the first-layer to calculate $I_X(\omega_i)$ is:

$$I_x(\omega_i) = \sum_{j=1}^{n} \alpha_j(1,2..20 \mid \omega) \times R_j + \sum_{m=1}^{2} \left\{ \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \beta_{j,k}(\omega) \times A_{j,m} \times A_{k,m} \right\} + C(\omega) \quad (2)$$

Here, subscripts j and k denote positions j and k in the window around residue i. R_j is a 20-D vector with the 20 normalized position specific scoring matrix (PSSM) value in position j, and $\alpha_j(1,2 \ldots 20|\omega)$ is the coefficient vector for the state ω. $\beta_{j,k}(\omega)$ is the coefficient of combining positions j and k. A is the value of various chemical and physical properties of the side chains, which was summarized previously,[10] and m denotes the respective properties of the side chains. $C(\omega)$ is a constant.

The first section of Eq. (2) represents the contribution of single residues in the window. The second section of Eq. (2)

represents the contribution of pair interactions of residues. Using a window size of 17, there are 136 ($17 \times 16/2$) position combinations and 399 amino acid pair combinations, with a total of 54,264 coefficients for the pair interactions. To make it possible to determine the solution in practice, we made the simplifying assumption that the various chemical and physical properties of residues are independent and contribute to the pairwise interactions. In previous studies, we used five chemical and physical properties of the side chains for secondary structure prediction: the residue mass, the free energy of transfer of the residue from oil to water, the charge state, the aromaticity and the ability to make a side chain hydrogen bond;[10] the first two of these properties were used for solvent-accessibility prediction.[11] We found that including the last three chemical and physical properties of the side chains in the secondary structure prediction only slightly increased the prediction accuracy (by 0.1%, data not shown), but it increased the number of parameters to be determined by about 408. Therefore, in this study, we only used the first two chemical and physical properties of the side chains in both the secondary structure and solvent-accessibility predictions. For a window size of 17, Eq. (2) consists of 340 (17×20) coefficients for the 20 amino acids, 272 (2×136) coefficients for the combination of residue mass and transfer free energy, giving a total of 612 parameters for every structural state to be determined.

All the parameters were determined using the data in a training set with the multiple linear regression method to minimize the sum of the square of deviations between the left and the right side of Eq. (2), as described previously.[10,11] The prediction was performed using a jackknife analysis by exploiting the protein sequence dataset, CB513. Each of the sample proteins in the dataset, in turn, was excluded from calculation of the coefficients. These coefficients were then used in Eq. (2) to calculate the information for residue i in state ωi, where $I_X(\omega_i)$ of the protein was excluded. The highest value of $I_X(\omega_i)$ was chosen as the prediction, that is to say, in a winner-take-all strategy.

The first-layer output structural state (H, E and C for secondary structure prediction, B and E for solvent accessibility prediction) was classified into five reliability groups according to the reliability index (see below). This result is an output of 15 possible secondary structural states (H_i, E_i, C_i; i = 1–5), and 10 solvent-accessibility states (B_i and E_i, i = 1–5). This information was used to build the secondary-layer prediction.

$$I_x(\omega_i) = \sum_{j=1}^{n} \alpha_j(1,2..20 \mid \omega) \times R_j + \sum_{m=1}^{2} \left\{ \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \beta_{j,k}(\omega) \times A_{j,m} \times A_{k,m} \right\} + \sum_{j=1}^{n} \gamma_j \times S_j + C(\omega) \quad (3)$$

Here γ_j is a coefficient vector for the predicted structural state in position j (14-D for secondary structure prediction and 9-D for solvent accessibility prediction, with one left out to guarantee that the solution will be a nonsingular matrix), and S_j is the first-layer predicted structural state in position j. The other parameters in Eq. (3) are as defined for Eq. (2). Figure 1 shows the dual-layer architecture of the PSIMLR system.

Reliability Index

The reliability index (RI) was defined as the difference between the value of the predicted state (with the highest value) and that of the next-most probable state. When $RI > 1.0$, this was set to $RI = 1.0$, so the values ranged from 0.0 to 1.0.

Assessment of Secondary Structure Prediction

The main parameter measuring the accuracy of the protein secondary structure prediction is the per-residue prediction accuracy, Q, which gives the percentage of all correctly predicted residues within the three-state (H, E, C) classes, and is almost always employed for assessment of prediction approaches.

A more rigorous measure (introduced by Matthews[21]) involves calculating the correlation coefficient for each target state. The correlation coefficient for state X (X = H, E, or C) can be calculated as follows:

$$C_x = \frac{TP_X TN_X - FP_X FN_{X_s}}{\sqrt{(TP_X + FP_X)(TP_X + FN_X)(TN_X + FP_X)(TN_X + FN_X)}} \quad (4)$$

where TP_X, TN_X, FP_X, and FN_X, are the true positives, true negatives, false positives and false negatives, respectively, of state X.

Both these previous methods concentrate mostly on prediction accuracy of secondary structural elements. In an effort to make evaluation of secondary structure prediction more structurally meaningful, the SOV (Segment Overlap) measure was proposed by Rost et al. (SOV94),[22] and redefined by Zemla et al. (SOV99).[23] Although the definitions of these two SOV measures are both exactly defined, the apparent accuracy obtained using these different definitions varies greatly. For example, the Jnet method has a SOV99 score of 74.2%, but a SOV94 score of 82.8% and the difference in the two scores for PHD is 9.8%.[17] Therefore, unless explicitly stated otherwise, the latest definition (SOV99) was used in this study.

Assessment of Solvent-Accessibility Prediction

Two methods were applied to assess the accuracy of the prediction: average Q and Matthews's correlation coefficient C_s. Q is a measure of the overall percentage of correctly predicted residues. The Matthews's correlation coefficient is the same as defined in the measurement of secondary structure prediction section (see above), but with replacement of the secondary structural states by solvent-accessibility states, as described previously.[11]

RESULTS AND DISCUSSION

Secondary Structure Prediction

The three-state secondary structure prediction was trained/tested on 513 proteins with a single omit jackknife

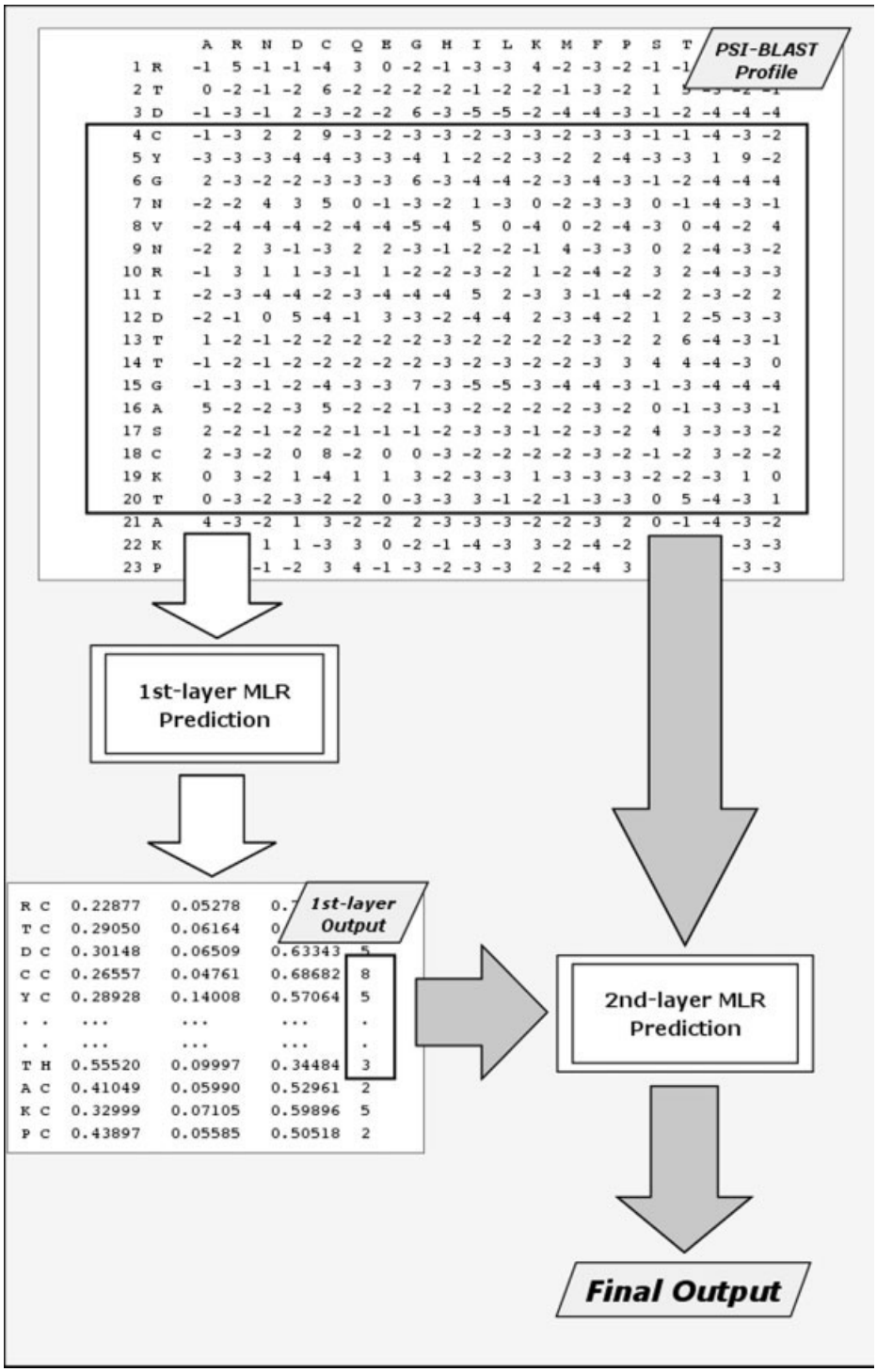

Fig. 1. The dual-layer architecture of the PSIMLR system. The input was the PSI-BLAST position specific profile. For each residue, a window size of 17 was used. The information used in the first-layer prediction is also exploited in the second-layer prediction. The reliability indices calculated from the first-layer output are additional information for second-layer prediction.

procedure, producing an average Q_3 accuracy of 76.4% and a SOV99 score of 73.2% (Table I, reduction Method A). Table I shows that the PSIMLR method is most successful in the prediction of coils (Q_C = 80.5%), quite good in the prediction of helix (Q_H = 79.1%), but less successful in specifying strands (Q_E = 64.7%). However, the trends indicated by the C_X values were different, with the highest value for C_H, and the lowest value for C_C. The cause may be the unequal frequencies of occurrence of different secondary structure elements in the test set.

TABLE I. The Prediction Accuracy of the Result Produced by the Full Dual-Layer PSIMLR System

	Predicted			
Observed	H	E	C	Total
H	23007	901	5189	29097
E	1109	12329	5621	19059
C	3741	3260	28962	35963
Total	27857	16490	39772	84119
Q_{obs} (%)	79.1	64.7	80.5	76.4
C_s	0.710	0.615	0.576	
SOV (%)				73.2

TABLE II. The Secondary Structure Prediction Accuracy of the First Layer Output

	Predicted			
Observed	H	E	C	Total
H	22,868	964	5265	29,097
E	1276	12,227	5556	19,059
C	3937	3190	28,836	35,963
Total	28,081	16,381	39657	84,119
Q_{obs} (%)	78.6	64.2	80.2	76.0
C_x	0.697	0.611	0.572	
SOV (%)				71.6

In theory, because the MLR method is based on the attempt to model the relationship between the independent variables and the dependent variable by fitting a linear equation to observed data, the advantage of the MLR method in comparison to neural network and machine learning-based approaches is that all parameters of the algorithm have direct physical meaning, and provide information about the relationship between sequence and structure. For example, in Eq. (2), parameters represent information about single residues and the pairwise interactions of those residues. Furthermore, in the MLR method the pair interactions of the residues were simplified by assuming that only two chemical and physical properties contribute, that is, the mass of the residue and the free energy of transfer of the residue from oil to water. This simplification clearly leads to a better result than considering the residue pair interactions in an integral manner. For example, when carrying out a prediction from target sequence alone (not using multiple sequence information), GOR IV, which considers the contributions of the single residues (referred to singlets) and the pairwise interactions of residues (referred to doublets), achieved a Q_3 score of 64.4%.[24] Under the same conditions, the MLR method achieved a Q_3 score of 69% using the simplified pairwise interactions of residues.[10] This result clearly demonstrates that the residue mass and hydrophobicity are the dominative properties contributing to secondary structure formation.

An additional advantage of the MLR method is its relative simplicity and low computational resource requirements. However, the method provides higher accuracy prediction than GOR, and even provided higher accuracy than some neural network and support vector machine based methods.

To facilitate comparison with other methods, we have shown the results for different reduction methods in Table III. A few trained/tested results of different methods have been performed on the CB513 set, which makes objective comparisons readily available. Using reduction method E, the GOR V algorithm achieved a Q_3 accuracy of 73.5% and a SOV99 accuracy of 70.8%.[19] Under this reduction scheme, the MLR method achieved a Q_3 accuracy of 79.3% and a SOV99 accuracy of 74.5%. Cuff and Barton reported a Q_3 accuracy of 75.2% for Jpred and of 76.9% for Jnet from sevenfold crossvalidated experiments on a 480 protein dataset derived from CB513.[17] Adopting the same reduction scheme that Cuff and Barton used, method B, the corresponding Q_3 accuracy for MLR is 77.5%. Although the dataset of 480 proteins is not the same as CB513, because the new dataset was derived from CB513 by removing potentially poorly predicting chains (length less than 30 or without valid PSI-BLAST alignment profiles), the comparisons are nevertheless valid. A comparable performance based on a support vector machine method is that of PMSVM,[25] with a reported Q_3 accuracy of 75.2%, which is lower than MLR. The exact number of crossvalidation folds is not important provided the test set is representative and comprehensive.[9] The change in the crossvalidation methods from sevenfold or 20-fold crossvalidation to full jackknife showed a difference of less than 0.1% in the Q_3 accuracy.

A direct comparison of the most successful secondary structure prediction method, PSIPRED, to our method is not readily available. The only published crossvalidation result of PSIPRED, based on a set of 183 chains, gave a Q_3 accuracy of 76.5%.[13] This is a similar level of performance as achieved by the MLR method.

The first-layer PSIMLR prediction results are shown in Table II. Although the Q_3 value was only improved by 0.4%, the SOV99 value improved by 1.6%. The results reflect that the second layer filtered some noise from the first layer and made the result more accurate.

The prediction accuracy of the residues from the first-layer PSIMLR prediction with higher RI values is better than those with lower RI values, as plotted in Figure 2. Figure 2 also illustrates the grouping strategy that was used to divide each secondary structural state into five groups as the next layer prediction input.

The Reliability of Prediction for a Particular Protein

The distribution of prediction accuracy is described by a Gaussian distribution with mean $\mu = 77.5$ and variance $\sigma = 13.0$ as shown in Figure 3. This variation is a disadvantage where the interest is in a particular protein. The reliability index calculated from the output of the second layer has an approximately linear relationship with the prediction accuracy, as shown in Figure 4(A). This could then provide a reasonable estimate for the prediction accuracy of a full protein chain, by calculating the sum for every residue in the chain. The estimated per-residue overall prediction accuracy for a particular protein is as follows:

$$P_i = A + B^*R_i \quad (5)$$

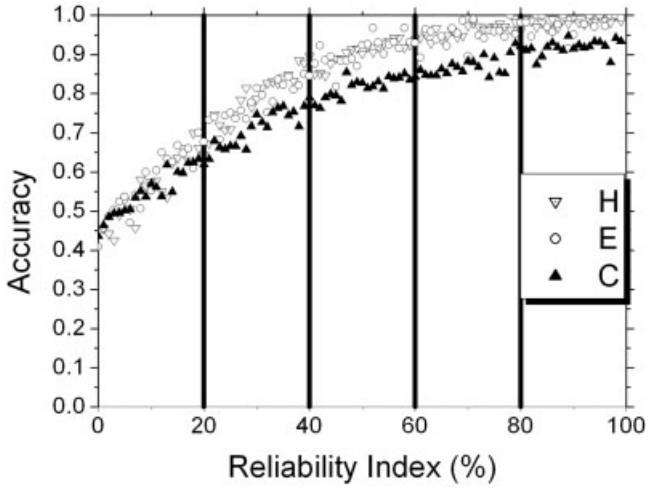

Fig. 2. The per-residue accuracy distribution of reliability indices of the first layer output, and the grouping strategy implemented for the second layer input. Each secondary structure element was divided into five groups at the 20th, 40th, 60th, and 80th percentiles.

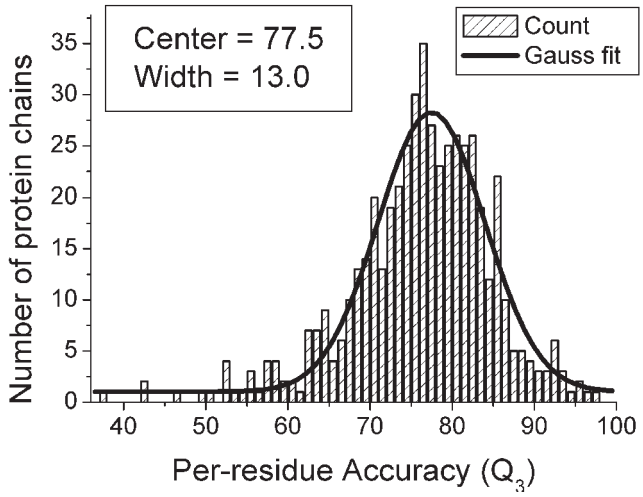

Fig. 3. Distribution of overall per-residue accuracy.

$$Q_{\text{est}} = \frac{1}{n}\sum_{i=1}^{n} P_i \qquad (6)$$

where n is the length of the protein chain, and P_i is the predicted accuracy of residues i calculated from the reliability index R_i. A and B are parameters extracted from the linear fit [Fig. 4(A)]. As shown in Figure 4(B), the estimated prediction accuracy gives a fairly good estimation of the performance of the prediction.

Prediction Accuracy Varies with Helix/Strand Length

We also examined the relationship between the prediction accuracy and the segment length of helices and strands. As shown in Figure 5, segments shorter than five for helix (HHHH or less) and shorter than three for strand

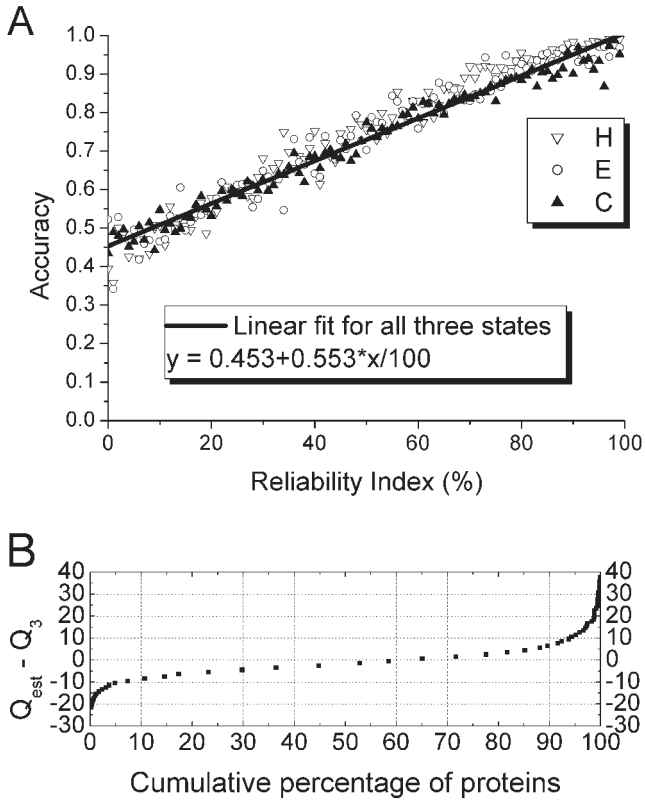

Fig. 4. (**A**) The per-residue accuracy distribution of reliability indices of the second layer output. The reliability index of a residue is linear relative to its prediction accuracy, as shown in the figure. (**B**) The variation between the estimated prediction accuracy and the true prediction accuracy. The Q_{est} provided a reasonably accurate estimate of the performance: for more than 90% of all proteins the function yielded estimates in the range $<\pm 10\%$ accuracy.

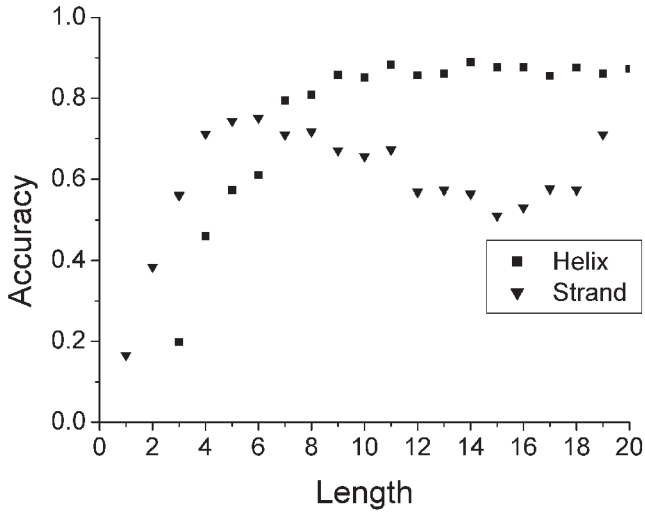

Fig. 5. The per-residue prediction accuracy varies with the segment length of helix and strand. Short segments (HHH and HHHH for helix, and E, EE for strand) were predicted poorly. Strands longer than fourteen also tended to have low prediction accuracy. However, because of the rareness of such segments (data not shown), the effect is small.

(EE or E) have a prediction accuracy of less than 50%, which means that the system has poor ability to discriminate between such segments and coils. Therefore, assign-

TABLE III. The Prediction Accuracy of Q and SOV Using Different Reduction Methods

Method	Q_3 (%)	Q_H (%)	Q_E (%)	Q_C (%)	SOV (%)
A	76.4	79.1	64.7	80.5	73.2
B	77.5	84.7	64.7	79.0	74.0
C	78.3	80.4	64.4	83.4	75.7
D	78.9	80.4	63.2	84.4	75.1
E	79.3	82.1	66.3	82.9	74.5

Method A: H, G, I to H; E, B to E; rest to C.
Method B: H to H; E, B to E; rest to C.
Method C: H to H; E, to E; rest to C.
Method D: GGGHHHH redefined as HHHHHHH, then B and GGG to coil, then H to H; E to E; rest to C.
Method E: H to H; E, to E; rest to C including EE and HHHH.

TABLE IV. Comparison of Average Prediction Accuracy and Matthew's Correlation Coefficient of the Solvent Accessibility Prediction

Prediction method	Average prediction accuracy (Q) (%)	Correlation coefficient (C_s)
MLRacc[b]	75	0.43
PSIMLRacc1[a]	77.2	0.537
PSIMLRacc2[a]	77.7	0.548
Jnet[c]	76.2	
BRNN[d]	77.49	

[a]PSIMLRacc1 and PSIMLRacc2 represent the output of the first layer and of the second layer, respectively.
[b]This result was reported in our previous study,[11] with a threshold of 20% for a two-state definition of solvent accessibility.
[c]This result was obtained from a dataset of 480 chains derived from CB513 through seven-fold crossvalidation experiments, with a threshold of 25%.[17]
[d]This result was obtained from a 1086 protein dataset with sequence identity of less than 22% through three-fold crossvalidation experiments, with a threshold of 25%.[26]

ing such segments as coil is reasonable, the result of which is similar to using reduction Method D or E. Given that the secondary structure prediction results will ultimately be used for fold recognition, this systematic error must be taken into account. This alteration may be compatible with further application of the prediction information. As shown in Table III, this adjustment will give slightly higher accuracies of prediction.

Solvent-Accessibility Prediction

This method was also developed to predict relative solvent accessibility. The jackknifed result is shown in Table IV. In this table, we also included results produced by other systems for reference. The PSIMLR based system ranked at the top of currently available solvent-accessibility prediction methods.

The dataset that Jnet[17] trained/tested is similar to CB513. Because the crossvalidation number has almost no effect on the prediction accuracy, as described above, the comparison is objective. The PSIMLR method performed 1.5% better. The dataset that BRNN[26] used is not readily available, so the direct comparison of PSIMLR and BRNN is not objective. However, the similar accuracy levels indicate that these two methods performed at similar level of accuracy.

Application and Future Improvement

Because the consensus of different high-accuracy prediction methods always continues to improve on the predictions of individual methods,[27] the improved MLR method represents a contribution to the protein structure prediction community.

Recent efforts have been made to predict the eight states assigned by DSSP[28] and to directly predict the backbone torsion angle.[29] Such attempts will substantially improve modeling procedures for local structures of protein sequence segments, and this is likely to be the direction of further development in secondary structure prediction. We will continue to endeavor to improve and simplify our algorithm, which should contribute to fold recognition methods that incorporate secondary structure prediction.

AVAILABILITY

The Web server that implements the improved multiple linear regression system for secondary structure and relative solvent accessibility prediction is accessible at http://spg.biosci.tsinghua.edu.cn/.

ACKNOWLEDGMENTS

We express our sincere thanks to Dr. Sarah Perrett of the Institute of Biophysics, Chinese Academy of Sciences, for comments on the manuscript.

REFERENCES

1. Norvell JC, Machalek AZ. Structural genomics programs at the US National Institute of General Medical Sciences. Nat Struct Biol 2000;7(Suppl):931.
2. Burley SK. An overview of structural genomics. Nat Struct Biol 2000;7(Suppl):932–934.
3. Terwilliger TC. Structural genomics in North America. Nat Struct Biol 2000;7(Suppl):935–939.
4. Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shirouzu M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y, Kyogoku Y, Miki K, Masui R, Kuramitsu S. Structural genomics projects in Japan. Nat Struct Biol 2000;7(Suppl):943–945.
5. Heinemann U. Structural genomics in Europe: slow start, strong finish? Nat Struct Biol 2000;7(Suppl):940–942.
6. Maggio ET, Ramnarayan K. Recent developments in computational proteomics. Trends Biotechnol 2001;19:266–272.
7. Venclovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins 2003;53(Suppl 6):585–595.
8. Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR, Elofsson A. CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 2003;53)Suppl 6):503–516.
9. Rost B, Sander C. Third generation prediction of secondary structure. In: Protein structure prediction: method and protocol. Clifton, NJ: Humana Press; 2000.
10. Pan XM. Multiple linear regression for protein secondary structure prediction. Proteins 2001;43:256–259.
11. Li X, Pan XM. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
13. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
14. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508–519.
15. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.

16. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. Proteins 2001;Suppl 5:192–199.
17. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
18. Salamov AA, Solovyev VV. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J Mol Biol 1995;247:11–15.
19. Kloczkowski A, Ting KL, Jernigan RL, Garnier J. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins 2002;49:154–166.
20. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 1973;79:351–371.
21. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
22. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
23. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.
24. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996;266:540–553.
25. Guo J, Chen H, Sun Z, Lin Y. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. Proteins 2004;54:738–743.
26. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
27. Ward JJ, McGuffin LJ, Buxton BF, Jones DT. Secondary structure prediction with support vector machines. Bioinformatics 2003;19:1650–1655.
28. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47:228–235.
29. Kuang R, Leslie CS, Yang AS. Protein backbone angle prediction with machine learning approaches. Bioinformatics 2004;20:1612–1621.