Multiple Linear Regression for Protein Secondary Structure Prediction

Xian-Ming Pan*
National Laboratory of Biomacromolecules, Institute of Biophysics, Academia Sinica, Beijing, China

ABSTRACT **In the present work, a novel method was proposed for prediction of secondary structure. Over a database of 396 proteins (CB396) with a three-state-defining secondary structure, this method with jackknife procedure achieved an accuracy of 68.8% and SOV score of 71.4% using single sequence and an accuracy of 73.7% and SOV score of 77.3% using multiple sequence alignments. Combination of this method with DSC, PHD, PREDATOR, and NNSSP gives Q_3 = 76.2% and SOV = 79.8%. Proteins 2001;43:256–259.** © **2001 Wiley-Liss, Inc.**

Key words: protein folding; secondary structure prediction; multiple linear regression; consensus; jackknife; amino acid sequence

INTRODUCTION

Prediction of the three-dimensional structure of a protein from its sequence is becoming a pressing problem for many biologists because the discrepancy continues to increase between the number of known protein sequences and the number of experimentally determined structures. The prediction of secondary structure from amino acid sequence is the most familiar and well-defined problem and is often regarded as the first step in understanding the protein-folding problem.

Secondary structure predictions have been performed by various methods.[1–17] Most of the currently successful secondary structure predictions take the nonlinear approach based on networks[7–9] and nearest-neighbor algorithms,[9,10] which are "black-box" predictors. They do not make the basis of their prediction explicit, nor do they provide insight into the principles governing the formation of secondary structure. Some methods take linear approach, such as DSC.[17] DSC applies GOR residue attributes, with the addition of hydrophobicity and amino acids position. In CASP3, the best method achieved prediction accuracy Q_3 = 74.6% and SOV = 73.6%.[1] Methods trained and tested on groups of aligned, homologous sequences are more accurate than that trained and tested on single sequence.[5–7,12] Prediction accuracy has also been improved by combining more than one algorithm.[5]

In the current work, a new multiple linear regression method was developed for protein secondary structure prediction. The prediction accuracy is 68.8% only using single sequence and 73.7% using multiple sequence alignments, respectively. Combining with DSC, NNSSP, PHD, and PREDATOR the accuracy is >76%.

MATERIALS AND METHODS

Database and Structure Definitions

The proteins used in this study were a set of 396 chains (CB396) representative of high-resolution structures with multiple sequence alignments compiled by Cuff and Barton.[5]. The secondary structural states were defined by DSSP program.[18] DSSP provides an eight-state assignment of secondary structure. There are four different published eight- to three-state reduction methods: method A: E and B to E, G and H to H, rest to C (coil); method B: E to E, H to H, rest to C; method C: E to E, H to H, rest to C including EE and HHHH; and method D: GGGHHHH redefined as HHHHHHH, then B and GGG to C, with H to H and E to E.

The values of Q_3 obtained by different reduction methods are about 3% different.[5] The complete database contains a total of 62,115 residues with a composition of 30% helix, 20% sheet, and 50% coil.[5] The prediction results of DSC, NNSSP, PHD, and PREDATOR were obtained from CB396 database. These results were obtained by using reduction method C, for comparing, in this work we mainly give numbers for using method C.

Algorithm

In a previous work, we applied the multiple linear regression method for solvent accessibility prediction.[19] Here, we are interested in predicting the secondary structural state of residue i, ω_i, based on knowledge of the amino acid sequence, $\{A_j\}$, of a "window" of restricted size n residues symmetric about location i. Locations within the window are indexed by j. Defining $I(\omega_i)$ as the structural information, the value of $I(\omega_i)$ was taken as 1 when the residue in position i being in the state ω_i, otherwise as 0 (e.g., if residue in position i being in helix state (H), then $I(\omega_H) = 1$, $I(\omega_E) = 0$ and $I(\omega_C) = 0$). According to Anfinsen's hypothesis,[20] $I(\omega_i)$, is determined by sequence of $\{A_j\}$:

$$I(\omega_i) = f\{A_i\} \tag{1}$$

Grant sponsor: National Natural Science Foundation of China; Grant number: 39625008; Grant sponsor: China Commission for Science and Technology; Grant number: 973 project G1999075606; Grant sponsor: 863 High-Technology Foundation: Grant number: 863-103-13-03-02.

*Correspondence to: X.-M. Pan, National Laboratory of Biomacromolecules, Institute of Biophysics, Academia Sinica, 15 Datun Road, Beijing 100101, China. E-mail: xmpan@sun5.ibp.ac.cn

Received 28 August 2000; Accepted 15 December 2000

TABLE I. Secondary Structure Prediction Accuracy Q_3 and Segment Overlap SOV

	Reduction method	Average accuracy Q_3 (%)	Segment overlap SOV (%)
Single sequence	A	66.5	68.2
Single sequence	B	67.4	69.1
Single sequence	C	68.8	71.4
Multiple sequence	C	73.7	77.3
Consensus	C	76.2	79.8

where $f\{A_i\}$ is an unknown function. We extend $f\{A_i\}$ to:

$$I(\omega_i) = \sum_{j=1}^{n} \alpha(1, 2 \ldots 19|\omega)_j R_j + \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \beta(\omega)_{j,k}\gamma(1, 2 \ldots 399) R_j R_k + C(\omega) \quad (2)$$

Here $\alpha(1,2\ldots19|\omega)$ is a coefficient vector of 19 amino acids (one left out) for state ω. R_j is a 19-D vector with the component for residue in position j as 1 and the others as 0 for prediction using single sequence and with aligned compositions of 19 amino acids in position j for prediction using multiple sequence alignments, respectively. $\beta(\omega)$ are coefficients of combining positions j and k.$\gamma(1,2\ldots399|\omega)$ is a coefficient vector of 399 pair combinations R_jR_k in positions j and k of 20 amino acids (one left out).

By using a window size of 21, there are 210 (21 × 20/2) position combinations and 399 amino acids pair combinations with a total of 83,790 coefficients for pair interaction. In practice, it is impossible to determine so many parameters, so Eq. 2 should be simplified. Assume that the interaction of residue pair is contributed by various interactions of chemical and physical properties of residue pair, and these interactions are independent. In this study, we assume that the residue properties of mass, free energy of transfer from oil to water, charge states, aromaticity, and ability to make side-chain hydrogen bond can contribute to the interaction of residue pair. So we have:

$$I(\omega_i) = \sum_{j=1}^{n} \alpha(1, 2 \ldots 19|\omega)_j R_j + \sum_{m=1}^{5} \left\{ \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \beta_m(\omega)_{j,k} A_{j,m} \times A_{k,m} \right\} + C(\omega) \quad (3)$$

Here, subscripts j, k, and m denote positions; j, k, and various chemical and physical properties of side chains, respectively. A contains values of various chemical and physical properties of side chains, which was summarized in Table III. $C(\omega)$ is a constant. For a window size of 21, Eq. 3 consists of 399 (21 × 19) coefficients for 19 amino acids, 5 × 210 coefficients for position combination of residue mass, transfer free energy, charge states, aromaticity, and ability of formation side chain hydrogen bond.

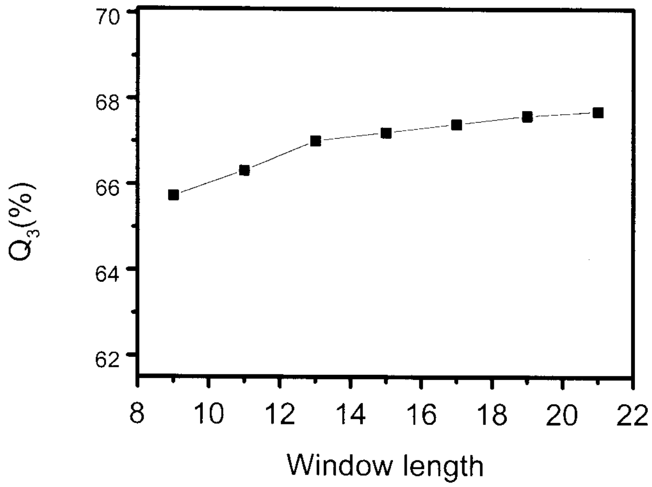

Fig. 1. Dependence of secondary structure prediction accuracy on window size.

TABLE II. Prediction Accuracy Varies With Helix/Strand Length

Helix/strand length	Helix % correct	Strand % correct
2	NA	22.3
3	NA	33.0
4	24.0	44.4
5	41.4	49.5
6	43.0	50.6
7	59.3	47.0
8	58.2	43.5
9	62.8	40.4
10	64.9	35.2
11+	73.2	34.9

All the coefficients were determined by the data in training set by using multiple linear regression method to minimize the sum of the square of deviations between the left and the right side of Eq. 3.[21] The prediction was performed with a "Jackknife" analysis exploited to the protein sequence database. Each of the sample proteins in the database, in turn, was excluded from calculation of coefficients. These coefficients were used in Eq. 3 to calculate the information of residue i in state ω_i, $I(\omega_i)$ of excluded protein. The highest value of $I(\omega_i)$ was chosen as the prediction.

The prediction results of PHD, DSC, NNSSP, PREDATOR, and this work were combined to make a consensus prediction. The consensus was calculated by examining the predictions for each method, at each position and taking the most popular state.

Accuracy Measures

All results reported were based on a window of length 21 residues, symmetric about the residue being predicted. To predict residue locations near the N and C terminals, 10 Gly were added to the ends of the chains, all taken in the coil states. Two methods were applied to assess the

TABLE III. Chemical and Physical Parameters of Amino Acids

Amino acids	A_1	A_2	A_3	A_4	A_5
G	0.0	0.0076	0.0	0.0	0.0
A	0.42	0.115	0.0	0.0	0.0
V	1.66	0.33	0.0	0.0	0.0
I	2.46	0.44	0.0	0.0	0.0
L	2.32	0.44	0.0	0.0	0.0
F	2.44	0.7	0.0	0.0	1.0
P	0.98	0.323	0.0	0.0	0.0
M	1.68	0.577	0.0	1.0	0.0
W	3.07	1.0	0.0	1.0	1.0
S	−0.05	0.238	0.0	1.0	0.0
T	0.35	0.346	0.0	1.0	0.0
N	−0.82	0.446	0.0	1.0	0.0
Q	−0.3	0.55	0.0	1.0	0.0
Y	1.31	0.82	0.0	1.0	1.0
H	0.18	0.63	1.0	1.0	1.0
D	−1.05	0.446	−1.0	1.0	0.0
E	−0.87	0.55	−1.0	1.0	0.0
K	−1.35	0.48	1.0	1.0	0.0
R	−1.37	0.777	1.0	1.0	0.0
C	1.34	0.36	0.0	1.0	0.0

A_1 = free energy of transfer from oil to water[23]; A_2 = relative side-chain mass with that of *W* as 1.0; A_3 = charge state of residue with positive as 1.0, negative as −1.0 and others as 0.0; A_4 = ability of formation side-chain hydrogen bond; A_5 = aromaticity with aromatic residues as 1.0 and others as 0.0.

accuracy of the prediction, average Q_3 and Sov. Q_3 is a measure of the overall percentage of correctly predicted residues. The other is the segment overlap[22]:

$$Sov = \frac{1}{N} \sum \frac{\min ov(s_{obs}{:}s_{pred}) + \delta}{\max ov(s_{obs}{:}s_{pred})} len(s_{obs}) \qquad (4)$$

where N is the total number of residues, min *ov* is the actual overlap, with max *ov* is the extent of the segment. $Len(s_{obs})$ is the segment length of observed structure. δ is the accepted variation that ensures a ratio of 1.0 where there are only minor deviations at the end of segments. δ = 1, 2, or 3 for segment length 1–5, 6–10, or longer than 11 and is restricted by:

$$\delta \leq \min\{\max ov - \min ov{:}\min ov{:}len(s_1)/2\} \qquad (5)$$

RESULTS AND DISCUSSION

Prediction Using Single Sequence

The three-state secondary structure predictions were performed only using single sequence over the 396 proteins with the single-omission jackknife procedure. This method achieved a Q_3 value of 66.5% and SOV score of 68.2% using reduction method A. The values of Q_3 and SOV obtained by using reduction methods B and C are summarized in Table I. Changing window size from 9 to 21 Q_3 increases from 65% to 67.4% (Fig. 1, using reduction method B). Because further increasing the window size is very time-consuming and only slightly improving the prediction accuracy, all predictions are performed on a window size of 21.

Prediction Accuracy Varies With Helix/Strand Length

The prediction results were sorted by actual length of the helix or strand; the percent correctly predicted residues in these elements are listed in Table II. Helix prediction accuracy increases with the length of the helix, whereas strand prediction accuracy peaks with length 5 or 6. Similar results have been also observed previously.[6] Short secondary structure elements (single turn helix and two E) are hard to predict, because they are normally marginal stable and variable within protein families.[4]

Prediction Using Multiple-Sequence Alignments

Prediction from a multiple alignment of protein sequences rather than a single sequence has long been recognized as a way to improve the accuracy.[5–7,12]. The three-state secondary structure predictions were performed from multiple alignment of protein sequences over the 396 proteins with the single-omission jackknife procedure. The average Q_3 accuracy is 73.7% and SOV score is 77.3% (Table I). There is ≈5% improvement in Q_3 accuracy and 8% in SOV score.

Consensus Prediction

The prediction results of ZPRED and MULPRED provided by CB396 database give obvious lower accuracy than other methods, so a consensus was calculated only from DSC, PHD, PREDATOR, NNSSP, and this work. Combination of these methods gives Q_3 = 76.2% and SOV = 79.8%, is higher than the consensus prediction without our method. Consensus prediction over the same database from DSC, PHD, PREDATOR, and NNSSP on CB396 gives Q_3 = 75.2%.[5] As reported by Cuff and Barton,[5] more complex methods for combining the different predictions give no further increasing in accuracy.

CONCLUSION

In the present work, a new method is developed for prediction of the protein secondary structure. The prediction method is simple, can be programmed easily, and can be extended easily to include new available information in prediction. The prediction accuracy of our method is comparable with the best existing methods. Similar to DSC our method is a linear approach, an advantage of which is that the prediction method is both implicit and effective.[5]

ACKNOWLEDGMENTS

We thank Professor C.L. Tsou of the Institute of Biophysics, Academia Sinica, for many helpful suggestions. The program has been written in C++ language and runs on a platform with Borland C++. It is free for the academia user and can be obtained by E-mail to xmpan@sun5.ibp.ac.cn with subject "for secondary structural prediction."

REFERENCES

1. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initia three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999;Suppl 3:149–170.

2. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:222–245.
3. Lim VI. Structural principles of the globular organization of protein chain: a stereochemical theory of globular protein secondary structure. J Mol Biol 1974;88:857–872.
4. Zhang X, Mesirov JP, Waltz DL. Hybrid system for protein secondary structure prediction. J Mol Biol 1992;225:1049–1063.
5. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508–519.
6. Wako H, Blundell TL. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. J Mol Biol 1994;238:693–708.
7. Chandonia J-M, Karplus M. New methods for accurate prediction of protein secondary structure. Proteins 1999;35:293–306.
8. Barton GJ. Protein secondary structure prediction. Curr Opin Struct Biol 1995;5:372–376.
9. Bohm G. New approaches in molecular structure prediction. Biophys Chem 1996;59:1–32.
10. Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. Proteins 1997;27:329–335.
11. Solovyev V, Salamov A. Local secondary structure prediction using local alignments. J Mol Biol 1997;263:31–36.
12. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 1993;90:7558–7562.
13. Biou V, Gilbrat JF, Robson B, Garnier J. Secondary structure prediction: combination of three different methods. Protein Eng 1995;2:185–191.
14. Nishikawa K, Ooi T. Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. Biochim Biophys Acta 1986;871:45–54.
15. Nishikawa K, Nogughi T. Predicting protein secondary structure based on amino acid sequence. Methods Enzymol 1995;202:31–44.
16. Geourjon C, Deleage G. Sopma: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci 1995;11:681–684.
17. Ross DK, Michael JES. Identification and applications of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 1996;5:2298–2310
18. Kabsch W, Sander C. Dictionary of protein secondary structures: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577–2637.
19. Xia Li, Xianming Pan. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
20. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
21. Zhang C-T, Zhang Z, He Z. Prediction of the secondary structure contents of globular proteins based on three structural classes. J Protein Chem 1998;17:261–272.
22. Rost BR, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
23. Michael JES. Protein structure prediction. New York; Oxford; 1996. 5 p.