Prediction of Protein Secondary Structure Content Using Amino Acid Composition and Evolutionary Information

Soyoung Lee, Byungchul Lee, Dongsup Kim*
Department of Biosystems, KAIST, Daejeon, Korea

*To whom correspondence should be addressed. E-mail: kds@kaist.ac.kr

Abstract

There have been many attempts to predict the secondary structure content of a protein from its primary sequence, which serves as the first step in a series of bioinformatics processes to gain knowledge of the structure and function of a protein. Most of them assumed that prediction relying on the information of the amino acid composition of a protein can be successful. Several approaches expanded the amount of information by including the pair amino acid composition of two adjacent residues. Recent methods achieved a remarkable improvement in prediction accuracy by using this expanded composition information. The overall average errors of two successful methods were 6.1% and 3.4%. This work was motivated by the observation that evolutionarily related proteins share the similar structure. After manipulating the values of the frequency matrix obtained by running PSI-BLAST, inputs of an artificial neural network were constructed by taking the ratio of the amino acid composition of the evolutionarily related proteins with a query protein to the background probability. Although we did not utilize the expanded composition information of amino acid pairs, we obtained the comparable accuracy, with the overall average error being 3.6%.

1. Introduction

Protein secondary structure content is the proportion of each secondary structure of a protein. Formally, it is defined as the ratio of the number of residues in a certain secondary structure to the number of total residues of a protein. In this work, we employed the conventional classification by DSSP [1], i.e., the eight secondary structure types: α-helix, β-strand, β-bridge, 3-turn helix, π-helix, hydrogen-bonded turn, bend, and random coil.

Knowing the secondary structure content of a protein is often the first step towards getting more detailed knowledge on its structure and function. However, the results of experiments had not been sufficiently accurate [7].

Among the early attempts to predict the

secondary structure content, notable prediction methods were the multiple linear regression approach [2] and the artificial neural network approach [3]. In these methods, it is assumed that the information of the amino acid composition, $\{P(A),\ P(C),\ P(D),\ ...,\ P(Y)\}$, where A, C, D, ..., and Y represent the single-letter codes of twenty amino acids, is enough to build a successful predictor.

Liu and Chou [4] expended the idea and introduced the concept of the coupled amino acid composition considering two adjacent residues to expand the amount of information. After computing 20 x 20 = 400 pair amino acid occurrence probabilities ranging from P(A|A) to P(Y|Y),

$$\begin{pmatrix} P(A\mid A), & P(B\mid A), & P(C\mid A), & & P(Y\mid A) \\ P(A\mid B), & P(B\mid B), & P(C\mid B), & \cdots & P(Y\mid B) \\ & \vdots & & \ddots & \vdots \\ & & & & \\ P(A\mid Y), & P(B\mid Y), & P(C\mid Y), & \cdots & P(Y\mid Y) \end{pmatrix}$$

On the other hand, Chou proposed the concept of the pair-coupled amino acid composition considering the two adjacent residues regardless of their order [5],

$$\begin{pmatrix} P(AA), & P(BA), & P(CA), & & P(YA) \\ & P(BB), & P(CB), & \cdots & P(YB) \\ & \vdots & & \ddots & \vdots \\ & & & \cdots & \\ & & & & P(YY) \end{pmatrix}$$

Recently, Cai *et al.* [6] developed an artificial neural network approach based on Chou's pair-coupled amino acid composition. They achieved a remarkable improvement in prediction accuracy.

Instead of expanding the number of parameters, we increased the amount of information of the amino acid composition by considering evolutionarily related proteins with a query protein. It is possible to expect better prediction accuracy with more information. Furthermore, there is another reason why we expect better accuracy than other predictors; we use the ratio of the amino acid compositions to the background probabilities, those in nature, while those were solely used as parameters in other methods.

2. Methods

Dataset

The same dataset used in Chou's and Cai's methods [6] were prepared; the training dataset consists of 244 proteins, of which no more than 35% had homology with one another and the test dataset of 202 proteins, of which no more than 35% had homology with the others, nor with those in the training dataset.

Algorithm

Our approach utilizes the fact that the evolutionarily related proteins share the similar structure [8]. Therefore, instead of predicting the secondary structure content of a query protein alone, it is possible to increase the prediction accuracy by predicting those of the evolutionarily related proteins all together.

In details, we calculate the composition of the twenty amino acids of the proteins that are evolutionarily related to a query protein, which can be easily done by simple numerical manipulation of the frequency matrix obtained by running PSI-BLAST [9]. In addition, instead of using the amino acid compositions themselves as the artificial neural network [10] input, we use the ratio of those of evolutionarily related proteins with a query protein to the background probabilities, which are obtained by counting the number of twenty amino acids of all the representative proteins in nature.

The computed results for 244 training dataset were applied to an artificial neural network for learning. After that, testing on 202 proteins was performed to estimate the prediction error.

The detailed procedure is as follows;

(1) Calculating the background probability of each amino acid(X), $P_0(X) = \sum_k \frac{N_k(X)}{L_k}$, where k represents one of 3352 proteins in FSSP, one of the most-widely used non-redundant protein structure databases, L_k is the length of k-th protein, and $N_k(X)$ is the number of X amino acid occurrences in the k-th protein.

(2) Getting the frequency matrix, $S_i(j,X)$, of the i-th protein among 244 proteins for training and 202 proteins for validation obtained by running PSI-BLAST with six iterations, where $S_i(j,X)$ represents the composition of X amino acid at the j-th position of the multiple sequence alignment of all the proteins that are evolutionarily related to the i-th protein.

(3) The occurrence of amino acid X of all the proteins that are evolutionarily related to the i-th protein is given by $P_i(X) = \frac{\sum_{j=1}^{L_i} S_i(j,X)}{L_i}$. To sum up, this equation should be understood as the summation of all possible occurrences of a specific amino acid at each position from the multiple sequence alignment.

(4) Calculating $\frac{P_i(X)}{P_0(X)}$ as input data of the artificial neural network with values from procedure (1) and (3).

(5) Applying an artificial neural network to the computed results of 244 proteins for training. The architecture of the neural network, as shown in Figure 1, consists of twenty input data from procedure (4), eighty hidden nodes, and eight output nodes that represent the desired contents of eight secondary structure types of a protein derived from the DSSP file of the protein.

(6) Finally, estimating the accuracy of the predictor by applying it to 202 test proteins.

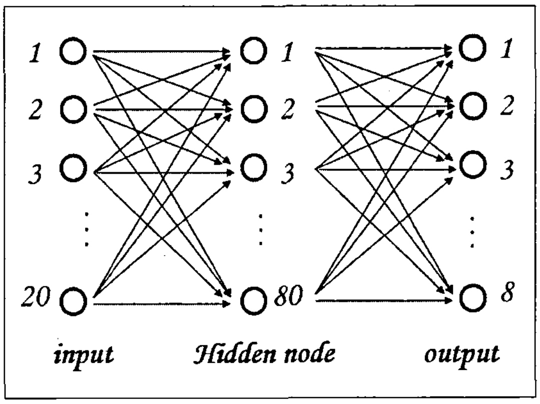

Figure 1. Artificial neural network architecture used in this work

Test criteria

There are two measures for the prediction error; the average absolute error and the overall average error. The first one is defined as $e^{\varphi} = \frac{\sum_{k=1}^{202} |\Theta_k - d_k|}{202}$, where e^{φ} is the average absolute error of φ structure, and k represents one of the test proteins. The predicted value and the desired composition of φ structure in k-th protein are denoted as Θ_k and d_k, respectively. Another test criterion is the overall average error, <e>. It is the mean of eight absolute errors, $< e >= \frac{\sum_{\varphi} e^{\varphi}}{8}$.

3. Results

The same tests were performed to compare our method with previous successful methods, Chou's and Cai's predictors [5, 6]. Thus, two types of tests were performed: a self-consistency test and an independent-dataset test.

Before testing error for the test dataset, the self-consistency test was performed by 244 training dataset to verify the consistency of training dataset and the fitness of learning. All of average absolute errors were less than 10% and the overall average error was 3.9%, indicating that our algorithm is consistent and the neural network was trained appropriately.

Consecutively, to confirm prediction accuracy, the independent-dataset test was tried by two test criterions: an average absolute error and an overall average error. In Figure 2, the average absolute error of our method for eight protein secondary structure types, H(α-helix), E(β-strand), B(β-bridge), G(3-turn helix), I(π-helix), T(hydrogen-bonded turn), S(bend) and C(coil) are compared with other methods, Chou's [5] and Cai's method [6].

The average absolute errors of our method are 0.085, 0.080, 0.0086, 0.022, 0.00056, 0.027 and 0.04 for H, E, B, G, I, T, S and C, respectively; most of them are less than other methods.

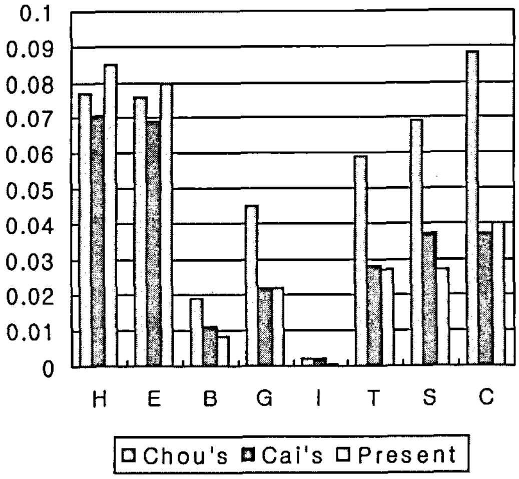

Figure 2. The average absolute errors of the present method are compared with those of Chou's and Cai's methods.

In the following table, the overall average error of our method is compared with that of the other methods.

Method	Chou's	Cai's	Present
Error (%)	6.1	3.4	3.6

The result shows that our method is better than that of Chou's method, while it is similar to that of Cai's method, the best accurate predictor until now.

In conclusion, the accuracy of our method, whose input vectors consist of 20 parameters, is comparable to the other accurate methods where input vectors consist of more than 200 parameters.

4. Discussion

The primary experimental method to determine secondary structure content is the circular dichroism (CD) spectroscopy [7]. However, it is well known that the accuracy of CD method is far from being satisfactory, with the error of roughly 10%. Surprisingly, the accuracy of predicting secondary structure content in fact exceeds that of experimental methods, with the error of roughly 3~4%.

In the previous works of secondary structure content prediction, the most accurate method, Cai's predictor, produced very successful result with the error of 3~4%. Instead of using the simple composition of twenty amino acids, they employed the pair-coupled amino acid composition, which included more than two hundred parameters.

Here we demonstrate that, when it is combined with the evolutionary information, the simple amino acid composition alone is informative enough to produce a predictor whose prediction accuracy is comparable to the most accurate predictors.

Most of average absolute errors of our method are less than Cai's method, the best predictor until now, while those of α-helix and β-strand are more than his method. We have to analyze the reason why his methods predicted the secondary structure contents of α-helix and β-strand successfully for the improvement of our method. By Liu and Chou's report [4], when they introduced the concept of coupled amino acid composition, the prediction errors were reduced up to about a half of the previous methods that considered simple amino acid composition. In other words, they demonstrated that the secondary structure content was related to the coupling effects of residues along a sequence which were not considered in our method. By this fact, we expect more accurate method when we consider the coupling effects of residues.

Therefore, we presume that a proper combination of the concept of pair-coupled amino acid composition and the evolutionary information might result in the more accurate predictor.

Acknowledgements

We thank the members of Protein Bio-Informatics Laboratory (PBIL), Sanjo Han, Seung Taek Ryu, and Chan-seok Jeong for helpful discussion. This work was supported by CHUNG Moon Soul Center for BioInformation and BioElectronics (CMSC).

Reference

[1] W. Kabsch, and C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637, 1983.

[2] W. R. Krigbaum and S. P. Knutton, Prediction of the amount of secondary structure in a globular protein from its amino acid composition, *Proceedings of the national academy of science of the USA*, **70**, 2809-2813, 1973.

[3] S. M. Muskal and S. H. Kim, Predicting protein secondary structure content: a tandem neural network approach, *Journal of molecular biology*, **225**, 713-727, 1992.

[4] W. Liu and K. C. Chou, Prediction of protein secondary structure content, *Protein engineering*, **12**, 1041-1050, 1999.

[5] K. C. Chou, Using pair-coupled amino acid

composition to predict protein secondary structure content, *Journal of protein chemistry*, **18**, 473-480, 1999.

[6] Y. Cai, X. Liu and K. C. Chou, Prediction of protein secondary structure content by artificial neural network, *Journal of computational chemistry*, **24**, 727-731, 2003.

[7] P. Pancoska, *et al.* Comparison of and limits of accuracy for statistical analyses of vibrational and electronic circular dichroism spectra in terms of correlations to and predictions of protein secondary structure, *Protein science*, **4**, 1384-1401, 1995.

[8] C. J. Epstein, Relation of protein evolution to tertiary structure, *Nature*, **203**, 1350-1352, 1964.

[9] S. F. Altschul, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25**, 3389-3402, 1997.

[10] S. Haykin, NEURAL NETWORKS: a comprehensive foundation, *Prentice Hall*, 2nd edition, 161-173, 1999.