Secondary Structure Prediction Based on Statistical Mechanics

Yukio Kobayashi [1] **Nobuhiko Saitô** [2]
koba@t.soka.ac.jp nsaito@mn.waseda.ac.jp

[1] Department of Information Systems Science, Faculty of Engineering, Soka University
1-236 Tangi-cho, Hachioji-shi, Tokyo 192-8577, Japan
[2] Department of Applied Physics, Waseda University, Shinjuku, Tokyo 169-8555, Japan

1 Introduction

Recent genome science has been developing from two sides: one is biological evolution due to mutation of DNA sequence and the other is biological function of protein molecules formed according to amino acid sequences. We have devoted ourselves to elucidate the mechanism of protein folding, and to develop the method for protein structure prediction [1, 2]. Our method has successfully yielded the native-like tertiary structures of various proteins, by using secondary structure information. Secondary structure prediction with a high precision, however, is essential to proteins of unknown structures. Our method for three-state (α-helix, β-strand and coil) prediction have reached the prediction accuracy of only 68%, which accuracy may not be sufficient for packing secondary structures into correct tertiary structure. Some methods [3] for secondary structure prediction exceed a little the accuracy achieved by our method. Contrary to the other methods, we can analyze the reasons for the poor accuracy and thus are attempting to improve the present accuracy. In this article, we describe the way to improve the prediction accuracy presented in Genome Informatics Workshop IV (GIW '93) [4].

2 Formulation

We formulated the statistical mechanical method for secondary structure prediction as follows: (1) The ith and the $(i+k)$th residues can interact with each other if and only if all of the ith to the $(i+k)$th residues exist in the same state (α-helix or β-strand). If even one residue in between is in a different state, the ith and the $(i+k)$th residues do not interact with each other. (2) A residue in coil does not interact with any other residues. (3) At least one residue in coil exists between neighboring secondary structures. (4) The distance of interaction k is taken up to 4 for α-helix and up to 2 for β-starnd. This is justified by considering the hydrogen bonds between the ith and the $(i+4)$th residues in α-helix and those between the ith and the $(i+2)$th residues in β-strand. In GIW'93 k was taken up to 4 also for β-strand.

We can obtain the probabilities $P(\alpha)_i$, $P(\beta)_i$ and $P(c)_i$ for the ith residue by calculating the partition function with recurrence relations [5]. Here, we need the statistical weights of $1620(= 20 + 20 \times 20 \times 4)$ residue pairs for α-helix and, in addition, $820(= 20 + 20 \times 20 \times 2)$ residue pairs for β-starnd. In contrast to GIW'93 the weights for $k = 3$ and 4 in β-strands were set at 1. To determine these weights, we optimized the objective function.

Table 1: Result of Prediction on 13 Proteins for Accuracy Estimation.

PDB-ID	No. of Residues	Accuracy (Previous) [4]	Accuracy (Present)
1ACX	108	58.333	62.037
1CTF	68	58.824	52.941
1LH1	153	85.621	84.967.
1UBQ	76	69.737	77.632
2ALP	198	60.606	51.010
2CDV	107	73.832	77.570
2CI2	65	86.154	58.462
2WRP	104	53.846	56.731
4RHV	255	62.238	65.882
3CLN	143	81.884	62.937
4FXN	138	61.314	89.855
5LYZ	129	56.589	64.341
7RSA	124	74.194	72.581
Total	1668	67.686	67.626

3 Results

We increased the number of proteins for optimization from 65 in [4] to 80 and additionally corrected the upper limit of the distance of the interactions from 4 to 2 between residues in β-strand as described in Section 2. We optimized the weights of 2440 ($= 1620 + 820$) pairs by referring to 80 proteins which do not have sequence homology among them, and then estimated the prediction accuracy for 13 proteins not included in, and not homologous with, the above 80 proteins. The accuracies were 67.6% for three states, 65.4% for α-helix and 45.9% for β-starnd.

4 Discussion

Now, we consider the reasons why the present accuracies are poor: (1) 265 pairs among 2440 ones are not found in the secondary structures of the 80 proteins for optimization. The weights of the missing pairs cannot be optimized properly. (2) The weights of 2175($= 2440 - 265$) pairs have not yet optimized sufficiently. The accuracies of prediction for the 80 proteins for optimization are only 78.3% in the present case. (3) An important problem is that the secondary structures, especially those around the edges of the structures are more or less deformed at the final stage of folding. Consequently, the secondary structure prediction cannot yield the correct results. In this connection we refer the proposal of Rost, Sander, and Schneider [6]. They showed that considerable variation in the position and length of secondary structure segments can be accommodated within the same three-dimensional structure. Thus, the goal of the precision accuracy can be reduced to some extent, and a new measure of segment overlap is introduced to compromise between permissiveness and precision.

The total number of the residues predicted correctly in the 80 referring proteins is increasing through the optimization, especially from 1527 residues to 1709 ones in β-strands. Thus we have a hope to improve the prediction accuracy by the present method.

Acknowledgements

This work has been supported by the Grants-in-Aid for Scientific Research No.10680644 from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Saitô, N., Kobayashi, Y., Ota, M., and Mitaku, S., Mechanism of protein folding and its application to structure prediction of proteins, *Rep. Prog. Polym. Phys. Jpn.*, 35:1–22, 1992.

[2] Kobayashi, Y., Sasabe, H., and Saitô, N., Ab initio method for predicting tertiary structures of globular proteins, *Fluid Phase Equilib.*, 144:403–413, 1998.

[3] Gromiha, M.M. and Selvaraj, S., Protein secondary structure prediction in different structural classes, *Protein Eng.*, 11:249–251, 1998.

[4] Kobayashi, Y. and Saitô, N., Prediction of structures of globular proteins, *Genome Informatics Series No.4*, 293–299, 1993.

[5] Saitô, N., Statistical mechanics of DNA and protein suitable for computer calculation, *Cell Biophys.*, 11:321–329, 1987.

[6] Rost, B., Sander, C., and Schneider, R., Redefining the goals of protein secondary structure prediction, *J. Mol. Biol.*, 235:13–26, 1994.