Beta-sheet Prediction Using Inter-strand Residue Pairs and Refinement with Hopfield Neural Network

Minoru Asogawa
C&C Research Laboratories, NEC
Miyamae, Miyazaki, Kawasaki Kanagawa 213 Japan
asogawa@csl.cl.nec.co.jp

Abstract

Many secondary prediction methods have been studied, but the prediction accuracy is still unsatisfactory, since β-sheet prediction is difficult. In this research, we gathered statistics of pairs of three residue sub-sequences in β-sheets, calculated propensities for them. When a sequence is given, all possible three residue sub-sequences are examined whether they form β-sheets. A shortcoming is that many false predictions are made. To exclude false predictions and improve the prediction, we employed a Hopfield neural network, in which the natural limitations on protein tertiary structure and preference of chemically stable long β-sheet are expressed in a form of energy functions. To clarify the prediction for heads and tails of β-sheets, special variables are introduced, which are similar to the line process proposed by Geman.

Introduction

Many secondary prediction methods use subsequences from 7 to 21 consecutive residues, and guess secondary structures of the center residues (Rost 93). These methods work well for α-helices, because one turn of an α-helix consists of 3.5 residues, thus 7 consecutive residues suffices to guess the secondary structure of the center residue. On the contrary, prediction for β-sheets are still difficult. In a β-sheet, lateral residues which are connected with hydrogen bonds are usually separated by more than 10 residues and distances between them are not constant. For this reason, predictions based on consecutive residues are not for a β-sheet itself, but for a strand. So strand prediction is necessary for β-sheet prediction, but is not sufficient.

A novel β-sheet prediction method

In the research presented here, I used a protein tertiary structure database (PDB) to gather statistics of pairs of three residue sub-sequences (will be abbreviated as **TRS**) in β-sheets, and calculated the propensities of **TRS** pairs (will be abbreviated as **pTRSP**). These propensities are used to guess whether two sub-sequences of a test sequence compose a **TRS** pair or not.

An advantage of this method is that it examines all possible residue combinations and finds almost all residues in β-sheets. In β-sheets, **TRS**s are packed tightly together, therefore information on six residues suffices to guess a **TRS** pair, like an α-helix is correctly predicted using 7 packed residues. A shortcoming of this method is that false predictions are also included, particularly those which do not arise in nature due to the limitations on a protein's tertiary structure. There are two kinds of erroneous guesses: the first erroneous guesses lie in parallel to correct β-sheets, and the second erroneous guesses appear randomly.

Improvement using a Hopfield neural network for a prediction result

In this research, a Hopfield neural network(Hopfield 86) is utilized for post-processing the result obtained by **pTRSP**. In a Hopfield neural network, the natural limitations on protein tertiary structure and preference of chemically stable long β-sheet are expressed in a form of energy functions.

Classification of TRS pair and constraints among them

To represent the natural limitations on protein tertiary structure, **TRS** pairs are classified in four types, as follows (see Fig. 1):

- **Ah**; a **TRS** pair which has one set of hydrogen bonds between the center residue pairs in an anti-parallel β-sheet.
- **An**; a **TRS** pair which has two sets of hydrogen bonds at each of the end residue pairs in an anti-parallel β-sheet.
- **Ph**; a **TRS** pair which has two hydrogen bonds between residue-**i** and residue-**x**, and residue-**i** and residue-**z**.
- **Pn**; a **TRS** pair which has two hydrogen bonds between residue-**h** and residue-**y**, and residue-**j** and residue-**y**.

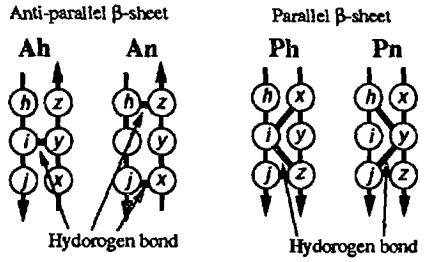

Figure 1: Four Types of **TRS** pair

This method is similar to the previous work (Hubbard 94) (Asogawa 96) (Krogh 96) (Mamitsuka 94). In this work hydrogen bonds patterns are well considered, which are essential for β-sheet prediction from both statistical and biological points of view (Asogawa).

Translation of the natural structure limitations to energy functions

A sequence with N residues is expressed with $(N-1)^2/2$ cells, each of with represents **TRS** pair. Since four connection types are necessary, $4 \times (N-1)^2/2$ cells are used in total. To clarify heads and tails of β-sheets and improve prediction, special variables are introduced for anti-parallel and parallel β-sheets. $2 \times (N-1)^2/2$ cells are used for this purpose.

$Ah_{i,j}$ is a real value of $[0.0, 1.0]$, representing that **TRS**$(\mathrm{r}_{i-1}, \mathrm{r}_i, \mathrm{r}_{i+1})$ and **TRS**$(\mathrm{r}_{j-1}, \mathrm{r}_j, \mathrm{r}_{j+1})$ is connected with **Ah**. (**TRS**$(\mathrm{r}_{i-1}, \mathrm{r}_i, \mathrm{r}_{i+1})$ is abbreviated as **TRS***i*, in the following description). Similarly, $An_{i,j}, Ph_{i,j}, Pn_{i,j}$ represent that **TRS***i* and **TRS***j* are connected with **An**, **Ph**, **Pn**, respectively. In an anti-parallel β-sheet, there is no distinction between the left and right strands, therefore following equivalences hold.

$$Ah_{i,j} \stackrel{\text{def}}{=} Ah_{j,i},$$
$$An_{i,j} \stackrel{\text{def}}{=} An_{j,i}.$$

Contrary, in a parallel β-sheet, there is a distinction between the left strand and the right strand, by considering figure 1, it is clear that following equivalence holds.

$$Ph_{i,j} \stackrel{\text{def}}{=} Pn_{j,i}.$$

By using $Ah_{i,j}, An_{i,j}, Ph_{i,j}, Pn_{i,j}$ both the natural limitations on protein tertiary structure and chemical stability are expressed as follows.

1. Each **TRS***i* can connect at most one **TRS** with **Ah**.
$$U_1 \stackrel{\text{def}}{=} \sum_j m(\sum_i Ah_{i,j} - 1).$$
$m(\cdot)$ is a function given as follows,
$$m(x) \stackrel{\text{def}}{=} \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise .} \end{cases}$$
The same equations hold for **An**,**Ph**,**Pn**.

2. The prediction should not diverge from the initial prediction value, which is nearly correct.
$$U_2 \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i,j}(Ah_{i,j} - Ah_{i,j}^{initial})^2.$$
$Ah_{i,j}^{initial}$ is the propensity of **TRS***i* and **TRS***j* having connected with **Ah**. The same equations hold for **An**,**Ph**,**Pn**.

3. When **TRS***i* connects with **Ah**, r_i uses all its hydrogen bond potential with **Ah**. Therefore **TRS***i* can not connect with **Ph**, which requires two hydrogen bond potential. In this case, **TRS***i* can connect with only **An** or **Pn**, which require no hydrogen bond potential.
$$U_3 \stackrel{\text{def}}{=} \sum_j m((\sum_i (Ah_{i,j} + Ph_{i,j}) - 1)).$$

4. Similarly, when **TRS***i* connects with **An**, both **TRS***i* -1 and **TRS***i* $+1$ use one hydrogen bond potential. Therefore **TRS***i* can not connect with **Pn**, which requires one hydrogen bond potential for both **TRS***i* -1 and **TRS***i* $+1$. In this case, **TRS***i* can connect with only **Ah** or **Ph**.
$$U_4 \stackrel{\text{def}}{=} \sum_j m((\sum_i (An_{i,j} + Pn_{i,j}) - 1)).$$

5. Each residue can have at most two lateral residues of any type.
$$U_5 \stackrel{\text{def}}{=} \sum m((\sum_i Ah_{i,j} + An_{i,j} + Ph_{i,j} + Pn_{i,j}) - 2).$$

6. For all **TRS** pair, at most one of **Ah**, **An**, **Ph** or **Pn** can be chosen.
$$U_6 \stackrel{\text{def}}{=} \sum_{i,j} m(Ah_{i,j} + An_{i,j} + Ph_{i,j} + Pn_{i,j} - 1).$$

7. In nature, anti-parallel β-sheets, **Ah** and **An**, line up alternatively. Therefore, when **TRS** pair$(i, j-1)$ of **Ah** adjoins **TRS** pair$(i-1, j)$ of **An**, an anti-parallel β-sheet gets longer and chemically stable. This condition is expressed as follows,
$$-\frac{1}{2}\sum_{i,j}\left(min\left\{(Ah_{i,j-1} - \frac{1}{2})(An_{i-1,j} - \frac{1}{2}),\right.\right.$$
$$\left.\left.(An_{i,j-1} - \frac{1}{2})(Ah_{i-1,j} - \frac{1}{2})\right\}\right).$$
However, this energy function tends to make longer prediction than the actual β-sheet and degrades prediction as a result. Therefore I used special variable SA to determine heads and tails of anti-parallel β-sheets. SA is closely related to the line process (Geman 84). When $SA_{i,j} = 1.0$, **TRS** pair$(i, j-1)$ and **TRS** pair$(i-1, j)$ is discontinuous, one **TRS** pair is in an anti-parallel β-sheet and the other is not.

When $SA_{i,j} = 0.0$, either both **TRS** pairs are in an anti-parallel β-sheet or both are not. By using SA, the chemical stability of an anti-parallel β-sheet is expressed as follows,

$$U_7 \stackrel{\text{def}}{=} -\frac{1}{2}\sum_{i,j}\left(min\left\{(Ah_{i,j-1}-\frac{1}{2})(An_{i-1,j}-\frac{1}{2}),\right.\right.$$
$$\left.\left.(An_{i,j-1}-\frac{1}{2})(Ah_{i-1,j}-\frac{1}{2})\right\}(SA_{i,j}-\frac{1}{2})\right).$$

8. Similarly, in natural parallel β-sheets, **Ph** and **Pn** line up alternatively. By using a special variable SP, which determines heads and tails of parallel β-sheets, the chemical stability is expressed as follows,

$$U_8 \stackrel{\text{def}}{=} -\frac{1}{2}\sum_{i,j}\left(min\left\{(Ph_{i,j}-\frac{1}{2})(Pn_{i-1,j-1}-\frac{1}{2}),\right.\right.$$
$$\left.\left.(Pn_{i,j}-\frac{1}{2})(Ph_{i-1,j-1}-\frac{1}{2})\right\}(PA_{i,j}-\frac{1}{2})\right).$$

9. In natural anti-parallel β-sheets, neither **Ah** nor **An** can line up continuously.

$$U_9 \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i,j}(Ah_{i-1,j+1}Ah_{i,j}+Ah_{i,j}Ah_{i+1,j-1})$$
$$+\frac{1}{2}\sum_{i,j}(An_{i-1,j+1}An_{i,j}+An_{i,j}An_{i+1,j-1}).$$

10. Similarly, in natural parallel β-sheets, neither **Ph** nor **Pn** can line up continuously.

$$U_{10} \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i,j}(Ph_{i-1,j+1}Ph_{i,j}+Ph_{i,j}Ph_{i+1,j-1})$$
$$+\frac{1}{2}\sum_{i,j}(Pn_{i-1,j+1}Pn_{i,j}+Pn_{i,j}Pn_{i+1,j-1}).$$

11. There are two other energies; one to limit the shortest length of β-sheet and one to lessen the number of heads and tails of of β-sheets.

An energy function of Hopfield neural network is given as a weighted sum of these energies U_k.

$$E \stackrel{\text{def}}{=} \sum_k \alpha_k U_k.$$

Appropriate αs are very important for Hopfield neural network work correctly. I roughly determined the αs and improved them using a learning method proposed by Kawato (Kawato 88).

Derivation of Hopfield neural network formula

To make a Hopfield neural network converge to a minimum of the energy function E, the steepest decent method is used. For this purpose, partial derivatives of E with respect to $Ah_{i,j}$, $An_{i,j}$,$Ph_{i,j}$ and $Pn_{i,j}$ are calculated. The derivative coefficient is used to update the membrane potential $mAh_{i,j}$ of the cell $Ah_{i,j}$.

$$\Delta mAh_{i,j} \stackrel{\text{def}}{\propto} -\frac{\partial E}{\partial Ah_{i,j}} = -\frac{\sum_k \alpha_k \partial U_k}{\partial Ah_{i,j}}.$$

Here, E is almost linear in $Ah_{i,j}, An_{i,j}, Ph_{i,j}$ and $Pn_{i,j}$, so the derivation of their partial derivatives is straight forward. The updated $mAh_{i,j}$ is used to determine $Ah_{i,j}$.

$$Ah_{i,j} \stackrel{\text{def}}{=} g(mAh_{i,j}) = \frac{1}{1+e^{-mAh_{i,j}}}$$

Similarly SP and SA are updated, following partial derivatives.

Experiment

Creation of pTRSP

To obtain **pTRSP**, the HSSP database is utilized, described in (Hubbard 94). The propensity for **TRS** pair $(r_1, r_2, r_3) - (r_4, r_5, r_6)$ of type **X** is calculated as follows,

$$\mathbf{pTRSP}((r_1,r_2,r_3),(r_4,r_5,r_6),\mathbf{X}) = -\log\frac{p((r_1,r_2,r_3),(r_4,r_5,r_6),\mathbf{X})}{p(r_1)p(r_2)p(r_3)p(r_4)p(r_5)p(r_6)}$$

where, $p(r_1)$ is the natural probability of r_1 appearing in the HSSP. Assuming the Boltzmann distribution between chemical energy and probability, a **pTRSP** corresponds to its chemical stability energy.

β-sheet prediction accuracy

256 test sequences are selected as test sequences. All test sequences have similarity less than 30% to all sequences used to calculate **pTRSP**. There are 89,941 residues; 23,152 residues are in α-helices, 22,704 residues are in β-sheets and 44,085 residues are in coils. As for a tentative result, before applying the hopfield neural network, $Q_{2,\beta}$ is 48.60% on average and SD(standard deviation) is 12.74%. After applying the hopfield neural network, $Q_{2,\beta}$ is improved as much as 87.68%(SD 7.90%) on average. This improvement is due to recognizing non-β-sheet residues correctly by applying the hopfield neural network. Actually, $Q_{\bar{\beta}}$ improves from 31.55% (SD 14.71%) to 88.61% (SD 8.45%). Since non-β-sheet residues are more than 77% of the test sequences, this improvement has substantial influence on $Q_{2,\beta}$ improvement. Although Q_β decreases from 96.74%(SD 9.46%) to 82.71%(SD 21.35%), it is still at a high level. Consequently, C_β improved, from 0.2724 to 0.6709.

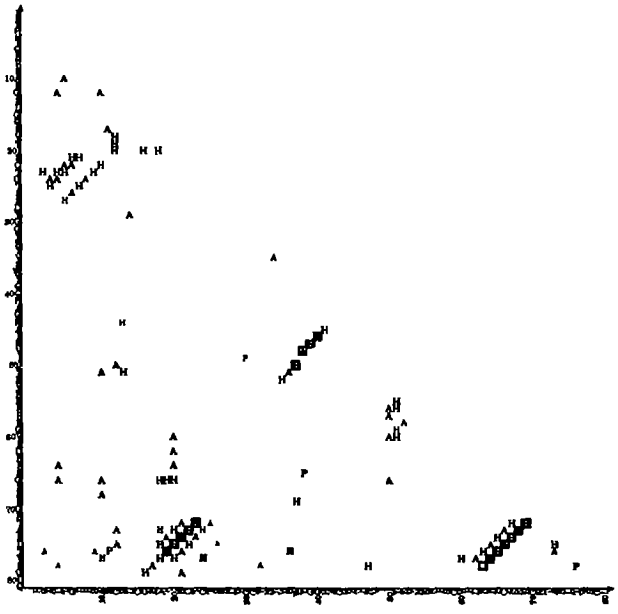

Figure 2: Prediction Result Before Applying a Hopfield Neural Network

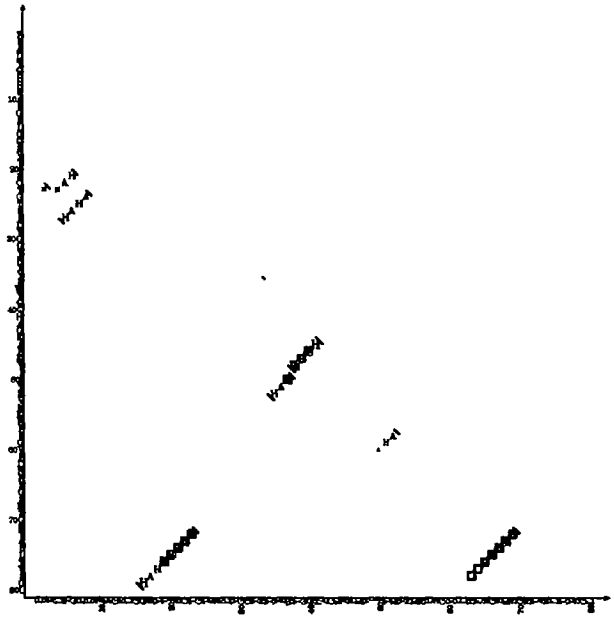

Figure 3: Prediction Result After Applying a Hopfield Neural Network

Part of the initial prediction and converged state is shown in figure 2 and 3, which are about **1mce** from first residue to 80th residue. In these figures, characters indicate predictions, rectangles indicate correct pairings, and slant lines indicate an activated SA. Note that where SAs are active, prediction for anti-parallel β-sheets are discontinuous and heads and tails of sheets are clearly defined.

Conclusion

In this research, the propensities of three residues pairs is used to predict β-sheets. This method finds almost all residues in a β-sheet if they are in it, and make many false predictions at the same time, however. To preclude those false predictions that are impossible due to the natural limitations on protein tertiary structure, I employed a Hopfield neural network to choose a prediction which satisfies tertiary structure limitations and which contains chemically stable long β-sheets. In the Hopfield neural network, the structure limitations and the preference of chemical stability are expressed in the form of energy functions. To clarify the prediction of the head and tail of a β-sheet, special variables are introduced, which are similar to the sheet processes proposed by Geman.

Acknowledgments This work is supported in part by the Ministry of Agriculture, Forestry and Fisheries of Japan in the contract of the neural network development for food manufacturing.

References

Asogawa M. and Fujiwara Y., "Beta-sheet Prediction Using Inter-strand Residue Pairs and Refinement Using a Hopfield Neural Network", *Abstracts of ISMB 96 Poster Presentation*, pp. 24, (1996).

Asogawa M., "Hydrogen Bonds Patterns are Essential for β-sheetprediction", ***unpublished.***

Geman S. and Geman D., "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Trans. on PAMI*, PAMI-6. vol. 6, pp. 721–741, (1984).

Krogh. A. and Riis S. K., "Prediction of beta sheets in proteins", *Advances in Neural Information Processing Systems*, vol. 8. (1996)

Hubbard T., "Use of β-strand Interaction Pseudo-Potentials in Protein Structure Prediction and Modeling", *Procd. of 27th Annual Hawaii International Conference on System Sciences*, pp. 336–344, (1994).

Hopfield J.J. and Tank D.W, "Computing with Neural Circuits: A Model", *Science*, (1986).

Kawato M., Ikeda T. and Miyake S., "Learning in Neural Networks for Visual Information Processing", *Journal of Television Society of Japan*, pp. 918–924, vol. 42, no. 9, (1988) (in Japanese).

H. Mamitsuka and N. Abe, "Predicting Location and Structure of Beta-Sheet Regions Using Stochastic Tree Grammars", *Proceedings of ISMB-94*, pp. 276–284, (1994).

Rost B. and Sander C., "Prediction of Protein Secondary Structure at Better than 70% Accuracy", *J. Mol. Biol.*, vol. 232, pp. 584–599, (1993).