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

Minoru Asogawa
asogawa@csl.cl.nec.co.jp

C&C Research Laboratories, NEC Miyamae, Miyazaki, Kawasaki Kanagawa 213 Japan

Over the last 20 years, many secondary prediction methods have been studied. Almost all methods use sequences with 7 to 21 consecutive residues, and guess the secondary structure of the center residue [Rost 93]. These methods work well for α-helices, because one turn of 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 is difficult. In a β-sheet, residues which are connected with hydrogen bonds are usually separated by more than 10 residues and the distance between them is not constant. For this reason, predictions based on consecutive residues are not for a β-sheet itself, but for a strand, which is a piece of stretched sub-sequence in the β-sheet. A β-sheet consists of a pair of strands, which are connected by hydrogen bonds. So a strand is a necessary for a β-sheet, but is not sufficient. In this research 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**) as described in [Hubbard 94].

[1] 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 a β-sheet if they are in it. In β-sheets, a **TRS** packed tightly with other **TRS**, therefore information on six residues suffices to guess a **TRS** pair, like an α-helix are correctly predicted using 7 packed residues. A shortcoming of this method is that false predictions are also included, which do not arise in nature due to the limitations on protein tertiary structure.

1 Improvement using a Hopfield neural network for a prediction result

An advantage of this method is that it examines all possible residue combinations and finds almost all residues in a β-sheet if they are in it. A shortcoming is that many false predictions are made. The nature of protein tertiary structure precludes the existence of these false predictions. To exclude false predictions and improve the prediction, I 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.

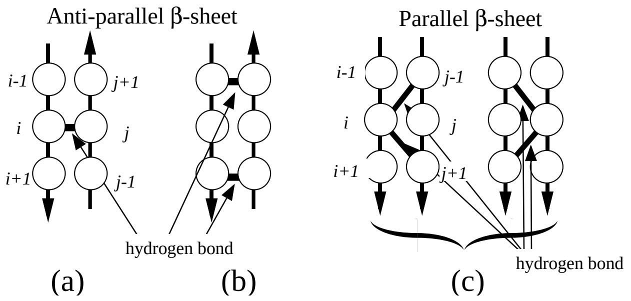
[1] We have utilized a layered neural network for **TRS** pair learning and used for secondary structure prediction. We could not surpass the prediction results obtained using **pTRSP**, however. The same attempt has been done independently by Krogh using 9 residue sub-sequences in β-sheets.

References

[Hubbard 94] 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).

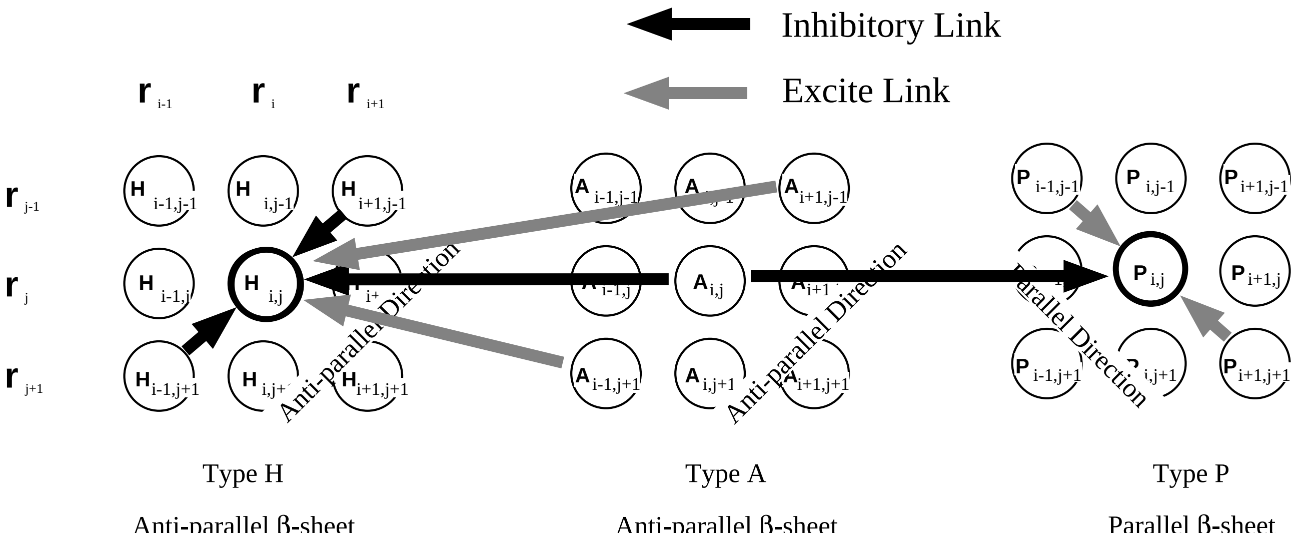
[Krogh 96] Krogh. A. and Riis S. K., "Prediction of beta sheets in proteins", *Advances in Neural Information Processing Systems*, vol. 8. (1996)

[Rost 93] Rost B. and Sander C., *J. Mol. Biol.*, vol. 232, pp. 584–599, (1993).



- (a) **Type H**; a **TRS** pair which has one set of hydrogen bonds between the center residue pairs in an anti-parallel β-sheet.
- (b) **Type A**; a **TRS** pair which has two sets of hydrogen bonds at each of the end residue pairs in an anti-parallel β-sheet.
- (c) **Type P**; a **TRS** pair in a parallel β-sheet.

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



There are three Hopfield neural networks; for **type H**, **type A** and **type P**. $H_{i,j}, A_{i,j}, P_{i,j}$ are the output activation values for cell (i, j), in each Hopfield neural network.

Figure 2: Hopfield neural network