A New Method for Predicting Long-range Interactions between Amino Acid Residues Based-on Homological Correlations

Hiroshi Mamitsuka
mami@sbl.cl.nec.co.jp

C&C Research Labs. NEC Corporation.
4-1-1, Miyazaki, Miyamae-ku, Kawasaki, Kanagawa, 216 Japan

Abstract

We propose a new method for predicting long-range interactions between amino acid residues based on what we term 'homological correlations.' Here two amino acid residues in a given sequence are said to be homologically correlated, if the substitution patterns of those positions in sequences homologous to the given sequence are correlated. Our method picks out a pair of amino acids of interest at a time, in general in distant positions, and predicts if they are homologically correlated.

An important characteristic of our method is that we enhance the input sequence(s) by obtaining sequences homologous to it, *not only* in training *but also* in testing. In particular, our method constructs a stochastic rule which takes as input the *changes* in the pair of positions of interest, and predicts whether or not there exists a long-range interaction between those positions, or more precisely it gives the likelihood for the pair to comprise a long-range interaction. On the basis of the likelihood calculated for each pair, our method finally predicts the pairs of positions comprising a strong long-range interaction, using a two-stage prediction method which consists of a type of heuristic-search algorithm and the Boltzman annealing technique[1].

In this paper, as a preliminary experiment for demonstrating effectiveness of our method, we focus on the problem of predicting the locations of disulfide bonds, which are a good example of long-range interactions. Disulfide bonds are covalent bonds which form between the side chains of two cysteine residues, adjacent in the three-dimensional structure, but located in distant positions in the primary sequence (e.g. [2]). Thus the problem of predicting the locations of disulfide bonds here is to determine the pairs of cysteines in a given sequence with unknown disulfide bonds, each of which forms a disulfide bond.

馬見塚 拓： NEC C&C 研究所, 〒216 川崎市宮前区宮崎 4-1-1

In our experiments, we extracted four proteins from the PDB_LIST 35% LIST[3] to meet a condition that at least 50 additional sequences which are homologous to each of four proteins are available from the HSSP (Homology derived secondary structure of proteins) database[4] Ver 1.0. The four proteins, each of which has less than 35% homology to the other three, are shown in Table 1. Our experimental result shows that, even when only one of the four proteins is used as training data, our method was able to predict *all* of the locations of disulfide bonds in all four proteins.

This result indicates that there exists a clear correlation between the substitutions of amino acids at any two positions which comprise a long-range interaction such as disulfide bonds. Also, this result suggests that our homological correlation based method is potentially useful in identifying various types of long-range interactions, such as helix-helix or helix-sheet contacts, any of which are thought to be crucial keys to predicting protein three-dimensional structures (e.g.[2]). At present, one biggest disadvantage of our homological correlation based method consists in use of a number of sequences for a given input in both learning and prediction, but such difficulty will be overcome in the future by immense increase of determined sequences with development of various kinds of genome sequencing projects, and then homological correlation will be greatly useful in predicting various long-range interactions described above.

HSSP code	# of disulfide bonds	# of obtained aligned sequences
1ppfe	4	53
1sgt	3	109
1ton	5	126
3rp2a	3	128

Table 1: Examples

References

[1] D. H. Ackley, G. E. Hinton, and T.J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[2] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc., New York and London, 1991.

[3] U. Hoboem, M. Scharf, R. Schneider, and C. Sander. Selection of a representative set of structures from the Brookhaven protein data bank. *Protein Science*, 1:409–417, 1992.

[4] C. Sander and R. Schneider. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.*, 9:56–68, 1991.