

An Evolutionary Approach to Finding Schemas for 3-Class Protein Secondary Structure Prediction

Huang, Hsiang Chi

ECRC-FIND, Institute for Information Industry

samuelhuang@ntu.edu.tw

Abstract. *A genetic algorithm has been applied to predict building schemas of protein secondary structure. This research uses protein secondary data generated by DSSP. Although the average Q3 of this research is not the highest score among previous researches, some fundamental and useful building schemas of protein secondary structure have been found. The results of this study would be a valuable reference for understanding the basic building patterns of protein secondary structures.*

1. Introduction

Assessing accurate secondary structures of a protein involves in preparation a crystal of protein, x-ray scanning and computing. These procedures cost a lot. Researchers have developed methods to predict secondary structures of a protein since 1960s. Recently, methods predicting protein secondary structure through the use of new algorithms such as HMM (3), neural networks (2), new evolutionary databases (3) etc.

These algorithms do help to predict protein secondary structure. However, some algorithms are like "Black Boxes". Researchers don't the meanings of understand enormous parameters or

how the results of prediction come out but only accept them. This study intends to predict protein secondary structure schemas by genetics algorithms. The results of this study would be a useful reference for understanding the basic building patterns of protein secondary structures.

2. Protein Secondary Structure Prediction by applying Genetic Algorithms

In past, several researches have been applied genetic algorithms to predict protein secondary structure. Some researches focused on global free energy minimum of protein secondary structure (8-12). However, these researches could not give us a complete understanding of the driving forces behind protein folding.

The second set of genetic algorithm methods describes another attempt to predict protein secondary structure by finding some useful building schemas of protein secondary structure using by genetic algorithm method (13).

This research intends to verify and bridge previous genetic algorithm researches. The first step of this research is finding the building schemas of protein secondary structure by applying Lin's genetic algorithm model (13).

Then classifying amino acid of building schemas is taken to figure out what is the driving force of protein folding.

3. Research Method and Materials

An implementation of a genetic system begins with encoding potential solutions to a specific problem on chromosome-like data structure. Previous researches (14-16) have proved that because of non-local interactions, information from more distant sequence could be improving prediction.

This research takes the first step by preparing protein sequences of $i \pm 4, 6, 8$ residues from DSSP database (17). The DSSP program classifies each residue into eight classes. These are typically reclassified into three standard classes with helices, strand, and random coil.

The structure of amino acid (i) of the sequence presents the schema of certain protein secondary structure (SS) as example below. (Chemical attribute) Each chromosome contains several schemas of all three type of protein secondary structure.

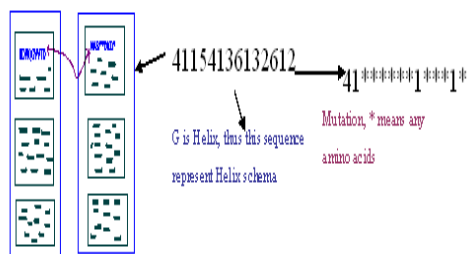


Figure 1. Data Structure

Then each schema is subcategorized by its R group into 7 families based on the chemical properties. In this research, crossover points are chosen at random in the population. This help us to find the better chromosomes (sets of schemas) Mutations occur in each point of a schema to simplify the rule of schemas. If there are more than schemas fitting to testing schema, then a vote mechanism is taken. Fitness functions are based on Q3. The modeling system of this study is listed as Figure2.

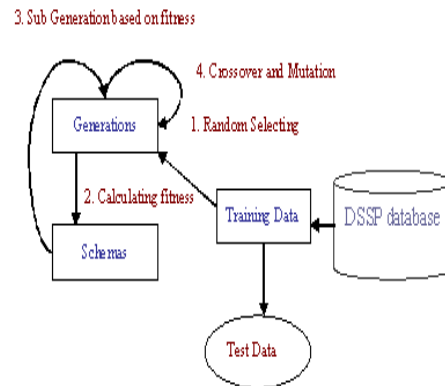


Figure 2. The modeling system

4. Results and Discussion

The Genetic algorithm system of this study were run at 50 populations, each population contain 80 schemas for each secondary structure (random coil, pleated sheet, helix) for 48 hours.

To verify schemas found by this GAPS (Genetics Algorithm for Protein Secondary Structure), a totally different protein data set is prepared for another tests. Prior studies have proved that if a total different protein data set is taken, Q3 achieved an average score about 46% based on different residues length

Although the highest Q3 of this research is not the highest score among researches, some fundamental and useful building schemas of protein secondary structure information have been found.

The results of GAPS (Genetics Algorithm for Protein Secondary Structure) contain schema of different orders. The word "order" of the schema refers to the number of amino acid that effects the construction of secondary structure of a protein.

With increasing the order of the sequence, the performance of Q3 rise and then falling significantly after order 4. The results have demonstrated that how many amino acids actually effect the conformation of protein secondary structure. From this study, order 1~4 of schemas achieved higher Q3 score.

Previous researches (e.g. focused on global free energy minimum of protein secondary structure) could not give us a complete understanding of the driving forces behind protein folding. Why?

Previous researches take every amino acid in the sequence into consideration. However, from the results of this study, not all the residues in a schemas effect the conformation of protein secondary structure. Only few amino acids actually effect the conformation of protein secondary structure folding.

5. Conclusion

Although this study demonstrated some interesting results, to state that there is a direct

correlation between the schema and protein secondary folding would be going too far. Some future works would be needed.

Recently, some researches have proved that combining different computing algorithm could improve prediction accuracy. (5) This study only applied genetic algorithms for protein secondary structure prediction. Combining with new computing technique (neural networks, HMM, multiple alignment) could be another way to improve prediction accuracy.

Acknowledge

Thanks are extend to Dr. Huang RenJung (Tamkang University), Dr. Dr. Lan-Yang Ch'ang (Taiwan Academic Sinica), Defong Liu and Shawn Lin (Institute for Information Industry) for their valuable advice and comments.

REFERENCES

- [1] Pauling, L. and Corey, R. B.(1951) Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets, *Proc. Natl. Acad. Sci. USA* 37,729-740
- [2] Jones, D.T.(1999) Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol.*292,195-202
- [3] Eddy, S. R.(1998) Profile hidden Markov models, *Bioinformatics* 14, 755-763
- [4] Baldi, P., Brunak, S. Frasconi, P., Soda, G., and Pollastri, G.(1999) Exploiting the past and the future in protein secondary structure prediction, *Bioinformatics* 15, 937-946.
- [5] Rost, B.(2001) Review: Protein Secondary Structure Prediction Continues to Rise, *J. Structural Biology*,0 ,1-15
- [6] Holland, J. (1975) *Adaptation In Natural and*

- Artificial Systems. University of Michigan Press.
- [7] Whitley, D. (1993) A Genetic Algorithm Tutorial. Technical Report CS-03-103, Colorado State University.
- [8] Unger, R., Moulton, J.(1993) Genetic algorithm for protein folding simulation. *J. Mol. Biol.*231, 75-81
- [9] Dandekar, T., Argos, P. (1993) Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.* 236,844–861
- [10] Pedersen, J.T., Moulton, J.(1995) Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms. *Proteins* 23:454–460
- [11] Cui, Y., Chen, R. S. and Wong, W. H. (1998) Protein Folding Simulation with Genetic Algorithm and Supersecondary Structure Constraints., *Proteins*, 31:247-257
- [12] McCammon, J. & Harvey, S. (1987). *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge.
- [13] Lin, F.K. and Sun, C.T. (2002) Protein secondary structure prediction using genetic algorithm. (C.S. Master thesis , Chinese Version) Natl. Chiao-Tung University, Taiwan
- [14] Levin, J., Robson, B. & Garnier, J. (1986).An algorithm for secondary structure determination in protein based on sequence similarity. *FEBS Lett*, 205,303-308
- [15] Yi, T.M. & Lander, E.S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.* 232,1117-1129
- [16] Qian, N. & Sejnowski, T. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* 202,865-884
- [17] Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983 Dec;22(12):2577