A System for Displaying Protein Structure Information based on Sequence Fragment Similarity

Shouji Tatsumoto

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2000

Keywords: proteins, sequence-structure interaction, sequence fragment, protein secondary structure, database.

Since the various reaction and phenomena in living things are carried by protein, science and engineering on protein have been eagerly promoted with great amount efforts. However, the ultimate principles which express protein folding and its functional expression are not known yet. In other words, today we can design and synthesize an amino acid sequence of a protein, but we can not know the structure and the function of it beforehand. Since the fact is preventing us from the systematic design and production of effective drags, It is very important to predict the structure and the function of a protein from its sequence information, i.e. string of 20 kinds of amino acid. The amount of the information about protein sequence and structure was very small. So, it was difficult to develop an algorithm for predicting protein structure from sequence in high accuracy. However, the situation has been drastically changed in these 10 years. Currently, about 27000 of protein sequences are gathered and publicly distributed by two major databases called PIR and SWISS-PROT, while about 11000 of protein structure are known and registered in a major database called PDB.

There are several methods for protein structure prediction which uses molecular energy calculations. However, most of them are not successful from the viewpoint of accuracy and computational complexity. On the other hand, approaches based

on database analysis like the threading method are rapidly improving their accuracy as the size of databases becomes large and large. In such a approaches, protein sequence homology plays an important role in understanding the relationship between structure and sequence. Furthermore, hybrid prediction methods emerged and thought as the most promising approaches, for example, threading and comprehensive homology search using PSI-BLAST. However, it still remains unsatisfiable level against the huge amount of protein sequences, which will be yielded by complete genome sequencing projects for model organisms. To devise an advanced and more accurate algorithm of structure prediction, it might be needed to scrutinize the sequence-structure correlation again from the viewpoint of short fragments.

To solve the problem described above, a system was developed in this study for displaying and analyzing ``how much sequence-structure correlation in the level of fragment are buried in the database of determined protein structures (PDB)".

The following are the procedures employed.

Step 1

When more than one same chain or high similarity chain exists in a large quantity, inclination occurs in the database. When a too huge size of database searches, its reference takes time too much. To normalize the biased distribution of PDB entries, PDB_SELECT was used.

Step 2

Representative protein chains were decomposed into short fragments with secondary structure information, and stored in a fragment database.

Step 3

A query sequence given by a user is compared with the fragment database, and the results are visualized in a web browser. It shows that secondary structure information drawable from normalized PDB entries is continuously fluctuating.

Results and Future work

We successfully attained accurate homologous sequence extraction and visualization of corresponding secondary structure. The prediction accuracy gets up to 74.9% for the whole of PDB. The prediction accuracy we obtained scores as high as the most accurate secondary structure methods. However, this system, still under intensive development, is not applicable to proteins which are less then 50 bases and more than 551 bases. And it only employs secondary structure information for prediction. Therefore, further integration of richer protein structural information, e.g. dihedral angel, surface or internal, and so on. Possible next strategy will be a hybrid approach such as applying existing methods for data which yield low predication accuracy with our system.

Publication

Tatsumoto, S. and Satou, K. "A System for Displaying Protein Structure Information Based on Sequence Fragment Similarity", Genome Informatics, 10, 237-238(1999)