PROTEIX: An Interactive Database System for Three Dimensional Protein Structures

Tatsuya Akitsu*
e-mail: akitsu@mel.go.jp

Mechanical Engineering Laboratory
1-2 Namiki, Tsukuba, Ibaraki 305 Japan

Abstract
We have been developing an interactive database system for three dimensional protein structures named PROTEIX. The most important feature of PROTEIX is that it has pattern matching facilities for three dimensional protein structures. For example, such functions as substructure search and alignment of two protein structures are included. As well as such functions, PROTEIX has the following facilities: a mechanism for inputting a file in PDB (Protein Data Bank) data format, an interactive graphic interface, and various pattern matching functions for amino acid sequences (strings).

Currently, PROTEIX is being developed and is used as a workbench for developing pattern matching algorithms. The final target of the system is to be a prototype of a powerful tool for biologists who study relations between amino acid sequences and three dimensional structures of proteins. PROTEIX is implemented on UNIX workstations using C-language. X-window is adopted for the graphic interface.

1 Introduction
A lot of database systems for analyzing DNA and amino acid sequences have been developed. However, a few database systems have been developed for analyzing three dimensional structures of proteins [3, 6, 7, 8] although three dimensional patterns have very important meanings. Thus, we have been developing a database system for analyzing three dimensional patterns of proteins, which is named PROTEIX (PROTEIn database system on uniX).

PROTEIX consists of three parts (see Figure 1): a data converter which converts a PDB format file into an internal data structure of PROTEIX, an interactive graphic interface for displaying three dimensional protein structures and inputting commands, and a set of pattern matching programs for three dimensional protein structures and amino acid sequences. All parts are written in C-language. X-window (X11R5) is adopted for the graphic interface. PROTEIX has high portability and it works on various UNIX workstations such as SUN SPARC workstations, SONY NWS-3460 workstations, HP 700-series workstations, and OMRON LUNA-88K workstations.

*阿久津 達也： 工業技術院 機械技術研究所（物理情報部 数理情報課），〒305 つくば市中里1-2
2 Pattern matching facilities

Currently, two new algorithms as well as conventional algorithms are implemented for fast substructure search and alignment of two proteins. Substructure search is, given a pattern structure $P$ and a text structure $Q$, to enumerate all parts of $Q$ similar to $P$. The alignment problem is, given two structures $P$ and $Q$, to find geometrically corresponding (equivalent) atoms between $P$ and $Q$. The method of the root mean square derivation [4, 5], which is a kind of least-squares fitting method, is used in all algorithms. Two new algorithms are named a least-squares hashing method and a dynamic matching method, respectively [1, 2].

The least-squares hashing method was developed for fast substructure search. It combines the hashing technique, which is a well-known technique in computer science, with the method of the root mean square derivation. The dynamic matching method was developed for alignment of two protein structures. It combines the dynamic programming technique with the method of the root mean square derivation. These two methods have been applied to PDB data. The results show that these methods have great advantages compared with a naive least-squares fitting method. For details, refer [1, 2].

Figure 1: Overview of PROTEIX
3 Graphic interface

PROTEIX has an interactive color graphic interface for displaying three dimensional protein structures and inputting commands. It has three modes (M1, M2 and M3) for displaying a three dimensional structure. In mode M1, all atoms and bonds are displayed. In mode M2, only carbon and nitrogen atoms in a backbone chain and bonds connecting them are displayed. In mode M3, an outline of the shape of a backbone chain is displayed as a curved tube. Moreover, in substructure search, matched parts of a text protein structure are highlighted. Of course, in each mode, structures can be rotated. It is possible to display a specified consecutive part of a protein only. This function is very useful to look at detailed substructures of proteins.

Figure 2: Snapshot of Graphic Interface
The graphic interface of PROTEIX is designed so that researchers who are not familiar with UNIX workstation can use it. Most commands are inputted using a 'mouse' device. Figure 2 illustrates how to input commands using a mouse device. Note that the system has a color graphic interface although it is not seen from Figure 2.

4 Conclusion

We have been developing PROTEIX. It is also used as a workbench for developing pattern matching algorithms for three dimensional protein structures. Since a lot of improvements have been done based on our experience, it has become a user-friendly tool.

However, development of PROTEIX is not yet completed. A lot of works should be done. Among them, the most important one is to make PROTEIX a 'true' database system. Since data are registered as text files in PDB format and they are converted into internal data structures when they are needed, such an operation as searching all protein structures with a specified substructure takes a long time. Thus, data structures (file structures) and methods which enable fast pattern matching for a large number of protein structures should be developed and implemented.

References


