3DinSight: An Integrated Database and Search Tool for Structure, Function and Property of Biomolecules

Jianghong An [1]
ajh@rtc.riken.go.jp
Takao Nakama [2]
nakama@rtc.riken.go.jp
Yasushi Kubota [2]
kubota@rtc.riken.go.jp
Akinori Sarai [1]
sarai@rtc.riken.go.jp

[1] Tsukuba Life Science Center, The Institute of Physical and Chemical Research(RIKEN),
3-1-1 Koyadai, Tsukuba, Ibaraki, 305 Japan

[2] Advanced Technology Institue Inc., 3-23-15 Jinbo, Kanda, Tokyo, 101 Japan

Abstract

We have created an integrated database, search and visualization tool, named 3DinSight, to help researchers to get insight into the relationship of structure, function and property of biomolecules. Various kinds of searches can be carried out though WWW interfaces. The locations of motif sequences and mutations are automatically mapped on the structure, and visualized in 3D space by interactive viewers, VRML (Virtual Reality Modeling Language) and RasMol. In the case of VRML, the mapped 3D objects are hyper-linked to the corresponding document data. The amino-acid properties of structure, functional and mutation sites, can be displayed as graph plots. 3DinSight is freely accessible through the ULR **http://www.rtc.riken.go.jp/3DinSight.html.**

1 Introduction

The structure, function and property of biomolecules are often closely related, but it is usually difficult to infer the relation from individual data. If researchers are interested in the structure of particular molecules and its relationship with function and physico-chemical properties, they usually need to examine several databases and literatures to obtain the information of their interest. It would be useful to have an integrated database where one can examine the relationship among structure, function and property. There are some services available in the Internet to link various databases, but the relational information of biomolecular structure, function and property is rather scarce. Thus, we have decided to create an integrated database and search/visualization tool to help researchers to get insight into their relationship.

2 Integrated database of structure, function and property

The coordinate data of PDB is used as the structural data. We have implemented the data into SYBASE relational database by creating various tables according to the major data items listed in the flat-format entry files of coordinate data. Those tables include document information such as source, authors and remarks, and structural information such as secondary structures and sequence.

As to the functional data, we have implemented PROSITE, which contains functional sites and motifs of proteins. The functional sites are usually described as sequence patterns. In order to make correspondence among the structural entry, the locations of functional sites on sequence and on structure, we have made pairwise alignments of the sequences in SEQRES and coordinate fields of PDB for all the protein chains by allowing gaps and insertions. Based on these alignments, we have created a table containing the pair of aligned sequences and sequence numbers. This table enables us to convert the amino-acid numbering in the SEQRES field to that in the coordinate field unambiguously.

Then, we have created a table for relating the protein chains with locations of functional sites in PROSITE.

We have also implemented the Protein Mutant Database(PMD), which contains a collection of information about protein mutants. The mutant positions are mapped on the amino-acid sequences of PDB by using the PDB-PIR cross-reference table. We created a table to relate the protein chains with locations of mutations to make one-to-one correspondence to the location on structure, in a similar manner to the functional site.

In order to relate structural and functional information with properties of molecules, we created a relational table containing various properties of amino acids based on their physico-chemical properties such as hydrophobicity and propensities for secondary structures. The table currently contains 43 representative properties of amino acids.

The present relational database also contains links to amino-acid databases, SWISS-PROT and PIR. Although these databases contain some cross-reference information, it is not necessarily complete nor updated. Thus, we decided to build cross-links between structural and sequence data in our relational database. For this purpose, we have compared the sequences in the SEQRES field for all the chains of PDB against those in SWISS-PROT and PIR, by using FASTA homology search program. Then, we selected only those entries with the highest homology.

3 WWW interface and visualization tool

We have built a WWW interface to the database so that researchers can access to the data and make various kinds of searches through the computer network. Two levels of interfaces are implemented. One is the form-based interface and the the other is the SQL-based interface. In the form-based interface, user can do most of regular searches such as entry, keywords, author and structure resolution simply by filling out the form. The sequence pattern matching allows ambiguities in sequence, length and repetition, and the patterns can be searched within particular secondary structures. The link to the functional data or mutant data can be made at the “Display Option” menu; e.g., if PROSITE link is selected, those structural entries with PROSITE functional sites will be displayed, Some motifs of PROSITE may occur in sequences by chance. Thus, the list also displays the frequency of the occurrence of each motif in PDB. We also prepared the interfaces for searching PROSITE, Protein Mutant Database, PIR and SWISS-PROT. The screened list can be linked to PDB entries, and displayed in a similar manner as before. Those users familiar with relational database can use the SQL-based interface. This interface provides powerful expression capability to search information under more complex conditions.

After the screening by whichever search method, each entry is linked to the original PDB files. From this entry screen, one can go to “Amino-acid analysis”, where properties of the molecule can be examined. Some mathematical operations such as sum and average for each protein chain can be performed and displayed. Furthermore, those properties can be plotted and displayed as a graph on the screen automatically. In this display, the actual locations of secondary structures, functional and mutation sites are automatically mapped, so that one can compare these locations with various property profiles of amino acids.

The screened structure and associated functional or mutation sites can be visualized by using visualization tools, VRML and RasMol. In either case, structures can be displayed automatically by selecting from the menu in the entry display. In the case of VRML, the functional or mutation sites in the 3D structure are clickable objects linked to the PROSITE or mutant documents, so that one can obtain the corresponding functional or mutational information such as description of the site and associated literature.