Construction of an Integrated Environment for Sequence, Structure, Property and Function Analysis of Proteins

Jianghong An [1] ajh@rtc.riken.go.jp
Takao Nakama [2] nakama@rtc.riken.go.jp
Yasushi Kubota [2] kubota@rtc.riken.go.jp
Hiroshi Wako [3] wako@mn.waseda.ac.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 Institute Inc., 3-23-15 Jinbo, Kanda, Tokyo 101, Japan.
[3] School of Social Science, Waseda University, Shinjuku-ku, Tokyo 169-8050, Japan

1 Introduction

One of the most important goals in molecular biology is to elucidate the relationship among sequence, structure, function and properties of biomolecules. Such knowledge would enable us to design the modifications of biomolecules for particular functions, and drugs to modify the function and property of biomolecules. Now, the number of entries in the Protein Data Bank (PDB) is over 10,000. These prized structural data should be used to understand the molecular mechanism of structural integrity and stability of biomolecules. The functionally important sites such as active sites in enzyme and ligand binding sites tend to be conserved among a family of proteins. The conserved amino acid sequences are called motif, and many motifs have been known so far. The physico-chemical properties of biomolecules are studied by various biophysical and biochemical methods. The structure, function and property of biomolecules are often closely related, but it is usually difficult to infer the relation from individual data. Thus, 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. On the other hand, structure comparison is one of the most important and interesting subject of bioinformatics because it plays a key role in structure classification, structure search, motif detection, function prediction, and so on.

It would be useful to have an integrated environment where one can examine the relationship among sequence, structure, function and property of biomolecules based on databases and a lot of search, analysis and visualization tools.

2 An Integrated Environment for Biomolecule Information

2.1 3DinSight: an integrated database

We have developed an integrated database of structure, function and property of biomolecules, called 3DinSight [1, 2], by focusing on the following points:

(1) Integrate PDB's structure, PROSITE's motif, PMD's Mutations and a lot of data of amino acid property into a relational database. The motifs and mutations are mapped to the 3D structures;
(2) Provide strong and flexible programs to search the database by keywords, sequences, motifs, and so on, which are very difficult for the original flat-formatted databases;
(3) Provide a World-Wide-Web (WWW) interface so that researchers around the world can access to the data and can carry out searches easily;

(4) Visualize the relationship among structure, functional sites and property automatically in 3D space, together with the link to associated document information.

2.2 Towards an integrated environment: new databases and tools

We are developing following new databases and tools, which are integrated to 3DinSight:

1) Thermodynamic Database for Proteins and Mutants(ProTherm)[3], which is a collection of various thermodynamic data such as Gibbs free energy, enthalpy, heat capacity, and so on of proteins and mutants. now over 5,000 entries are loaded in the database. This database and search tool will help researchers in studying the mechanism of stability of proteins and mutants.
2) Protein-Nucleic Acid Recognition Database, it consists of

   2.1) Protein-Nucleic Acid Complex Database, which is a collection of structural data of protein-nucleic acid complex. The database enables users to examine sequence-dependent DNA conformation in a form of table or graph. We plan to implement information on the conformation changes of protein upon complexation, and create interface to extract detailed information about base-amino acid interactions from the complex structure.
   2.2) Database of Base-Amino Acid Interactions, which collects pairs of atoms between bases and amino acids within 4 angstrom into a database table. User can specify residue names (base and amino acid), atom types (they can be checked by clicking "Atom Name") and side-chain/backbone to search the database. After the search, all the atom pairs with distance values will be displayed and all the atom pairs will be highlighted in the complex structure if you want to show the image. Thus, users can examine the specific interactions between base and amino acids in each structure.
   2.3) Thermodynamic Database for Protein-Nucleic Acid Interactions, which collects various thermodynamic data on interaction between proteins and nucleic acids;
   2.4) Tools for the predictions of binding sites and target genes of transcription factors.

   These databases and search tool will provide users with insight into the mechanism of protein-nucleic acid recognition from various aspects.
3) Protein-Ligand Database, which collects all ligands and the binding information with proteins from PDB. The ligand can be search by name, formula, structure and binding conditions.
4) Structure analysis tools of proteins, which are based on a novel method called Delaunay tessellation[4]. The interior space of the protein can be uniquely divided into Delaunay tetrahedra whose vertices are the $C\alpha$ atom positions. Then one unique code can be assigned to each tetrahedra by the vertex residues and four surrounding tetrahedron. Because the structure is represented in a string of digits, more easily and rapidly programs of structure analysis tools can be developed. The tools include structure classification, 3D structure searching of proteins, motif detection, and so on.

References

[1] An, J., Nakama, T., Kubota, Y., and Sarai, A., 3DinSight: an integrated relational database and search tool for structure, function and property of biomolecules, *Bioinformatics*, 14:188–195, 1998.

[2] Nakama, T., An, J., Kubota, Y., and Sarai, A., Visualization of functional sites on protein structures by virtual reality modeling language, *Bioimages*, 5:59–64, 1997.

[3] Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H., and Sarai, A., ProTherm: thermodynamic database for proteins and mutants, *Nucleic Acids Res.*, 27:286–288, 1999.

[4] Wako, H. and Yamato, T., Novel method to detect a motif of local structures in different protein conformations, *Protein Eng.*, 11:981–990, 1998.