PROTEIN INFORMATICS TOWARDS INTEGRATION OF DATA GRID AND COMPUTING GRID

HARUKI NAKAMURA
Institute for Protein Research, Osaka University
Osaka 565-0871, JAPAN
harukin@protein.osaka-u.ac.jp;
http://www.protein.osaka-u.ac.jp/rcsfp/pi; http://www.pdbj.org

Information of the structures and functions of protein molecules and their mutual interactions that construct protein networks increases rapidly as the consequence of the structural genomics and structural proteomics projects [1]. Advanced applications of such information require the Grid technology to solve the two problems: (i) the shortage of computational power, and (ii) the lack of a capability for seamlessly and quickly retrieving data from the varieties of heterogeneous biological databases [2].

1 Protein Structural Databases for Data Grid

Decades of researches in structural biology have already accumulated a large amount of structural data in PDB, and the advance of structural genomics is expected to increase the amount even more drastically. Our laboratory develops an advanced database for Protein Data Bank Japan (PDBj), collaborating with Research Collaboratory for Structural Bioinformatics (RCSB) in USA, and Macromolecular Structure Database at the European Bioinformatics Institute (MSD-EBI) in EU, in the framework of wwPDB [3]. We collect, curate, and edit the deposited coordinates of proteins mainly from Asian and Oceania countries.

A new extensible mark-up language (XML) describing the PDB data, the pdbML, is being developed by wwPDB [4]. Its structure is defined in XML Schema (pdbx-v1.000.xsd at `http://deposit.pdb.org/pdbML`), based on Macromolecular Crystallographic Information Format (mmCIF). The entire content in the pdbML is now available from `ftp://beta.rcsb.org/pub/pdb/uniformity/data/XML`. To make the most of the XML format, we, PDBj, have constructed an XML-based PDB data browser (xPSSS: xml-based Protein Structure Search Service at `http://www.pdbj.org/xpsss/`), using the native XML-DB. In addition to simple searches, full XPath searches are also implemented. This allows users to perform complicated searches, so that one can combine any searches on any items in the database. It also lets users control the output of their search in details. The xPSSS can be used by the SOAP service for large-scale analyses and data grid applications.

Beside the conversion to pdbML and its native database, which is a matter of formats, we are also trying to improve the content of the database. Although some PDB entries have extra information, in addition to the coordinates, such as experimental data and function information, most of them lack at least part of such information. We have been, therefore, complementing such missing data in the database, as completeness of data is often important to systematic analysis. Some items are incorporated from other database semi-automatically while others are extracted from literatures manually by annotators. Such additional information is also browsed in the xPSSS.

In addition, we have developed several secondary databases. Among them, a database, *e*F-site (electrostatic-molecular surface of Functional site: http://www.pdbj.org/eF-site), summarizes the protein molecular surfaces with their biochemical function information [5]. An algorithm using the clique detection method as an applied graph theory was developed for search of the *e*F-site database, so as to recognize and discriminate the characteristic molecular surfaces of the proteins. The method identifies the active site having the similar function to those of the known proteins, and could be used for search of the complementary surfaces, in order to analyze the protein-ligand and the protein-DNA interactions [6, 7].

A Java/OpenGL based viewer, PDBjViewer (jV), has been developed for the visualization of macromolecular structures as an applet and a stand-alone usage [5]. It can, in addition to various representations of atomic models, display arbitrary objects defined by vertices, lines, triangles or quadstrips, described by a simple XML format, so that the molecular surface and electron densities are easily visualized simultaneously with the atomic models. This viewer is also useful as a molecular graphics monitor in the grid architecture. The source codes of the program jV are available from our web site (http://www.pdbj.org/PDBjViewer/).

The protein structural alignment is performed by the ASH service (http://timpani.genome.ad.jp/~ash), with maximizing the number of equivalent residues (NER) to give us the correct structural neighbours of the individual proteins [8]. A new tool, Sequence Navigator, is now available at PDBj to search all structures having the homologous sequences, and the alignments are given back. Structure Navigator, in which similar 3D structures are searched with the NER algorithm, will start its service soon.

2 Computing Grid: Integration of Multiscale and Multilevel Simulations

In multiscale biological systems, integration of the computational methods for models at different levels is essential. A new platform, *BioPfuga* (Biosimulation Platform United on Grid Architecture), has been proposed and developed [2], where individual applications at different levels are united and executed as a hybrid application. *BioPfuga* requires that (i) application programs are divided into a set of many pieces, each of which corresponds to a unit simulation procedure, and that (ii) data communication be made among the program components by a standard XML description. For the first request, the program modules were implemented with a Grid service, GT3.2, defined by OGSI (Open Grid Services Infrastructure). An adaptor was useful to separate the grid middleware from the parallel computation made individually at local sites. For the second issue, a new XML description was designed for exchanging the information among program modules at different levels in three forms: text, hexadecimal, and Base64 forms. In particular, the data size with the Base64 form amounts only to 1.3 times of that with the binary form. The schema and API tools for the XML description are being fixed, and they will soon be available from our web site.

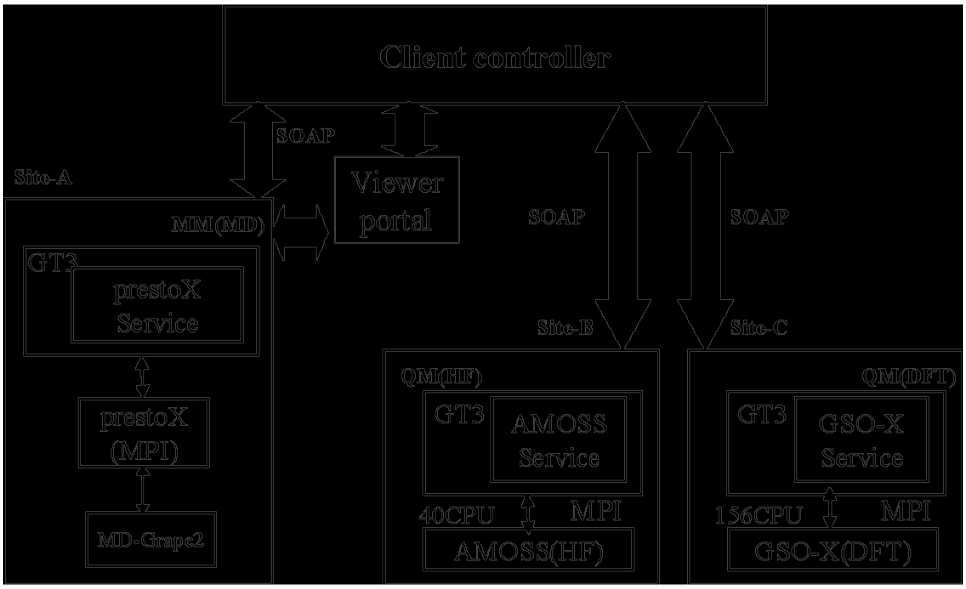

Figure 1. *BioPfuga* application system for hybrid-QM(HF)/QM(DFT)/MM method.

As an example of the *BioPfuga* application, we combined the quantum mechanical (QM) simulations with the program *AMOSS* for Hartree-Fock Molecular Orbital calculation (HF) developed by the NEC quantum chemistry group [9], and *GSO-X* for generalized spin density function theory (DFT) developed by Prof. Kizashi Yamaguchi [10], with the molecular mechanics (MM) simulations, the program *prestoX* [11]. The program *prestoX* has been developed to analyze the free energy landscapes of proteins, by developing our own algorithms of multicanonical molecular dynamics and Tsallis dynamics [12-14]. It was applied to predicting peptide structures, modelling loops, and flexible docking between a protein and the ligand [11, 15].

We first divided these three large programs into a set of component programs. Then, as shown in Figure 1, a hybrid calculation was performed using the PC clusters located in the very separated laboratories with their firewalls, and the MD part was also driven on the special-purpose computer for MD simulations, MDGrape2 [16]. An example of the *BioPfuga* application to hybrid-QM(HF)/QM(DFT)/MM method will be shown.

References:

1. Kinoshita & Nakamura. *Curr. Opin. Struct. Biol.*, 13:396—400, 2003.
2. Nakamura et al. *New Generat. Comput.*, 22:157—166, 2004.
3. Berman et al. *Nature Struct. Biol.*, 10:980, 2003.
4. Westbrook et al. *Bioinformatics*, 2004, in press.
5. Kinoshita & Nakamura. *Bioinformatics*, 20:1329—1330, 2004.
6. Kinoshita & Nakamura. *Protein Science*, 12:1589—1595, 2003.
7. Tsuchiya et al. *Proteins*, 55:885—894, 2004.
8. Standley et al. *Proteins*, 2004, in press.
9. Sakuma et al. *Int. J. Quant. Chem.,* 61:137—151, 1997.
10. Yamanaka et al. *Int. J. Quant. Chem.,* 91:376—383, 2003.
11. Fukunishi et al. *J. Phys. Chem. B* 107:13201—13210, 2003.
12. Kamiya et al. *Protein Science,* 11:2297—2307, 2002.
13. Fukuda & Nakamura. *Phys. Rev., E* 65:26105, 2002.
14. Kim et al. *Phys. Rev., E* 68:21110, 2003.
15. Watanabe et al. *J. Mol. Graph. Model.,* 23:59—68, 2004.
16. Narumi et al. *Mol. Simul.,* 21:401—415, 1999.