BioGrid: Integration of Biological Data Grid and Computing Grid

Haruki Nakamura

Laboratory of Protein Informatics

Research Center for Structural and Functional Proteomics

Institute for Protein Research

Osaka University

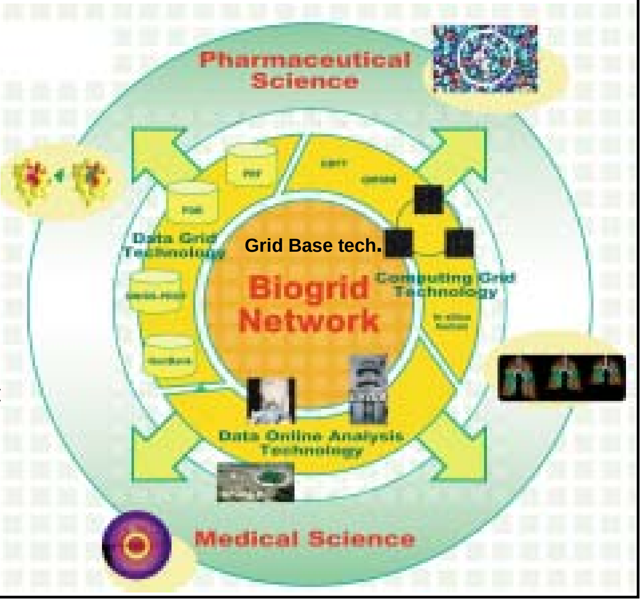
http://www.protein.osaka-u.ac.jp/rcsfp/pi/

BioGrid

http://www.biogrid.jp

- Started from 2002
- Goals

  Technology Development for: Biotechnology (drug discovery) and Medical Sciences.
- Leader: Shinji Shimojo (CMC, Osaka Univ.)
- Government Suppc (MEXT): 5years (3~4M$/year)



Project Organization

- **Grid Base Technology**: Establishing A Comprehensive System for Stable Operation Based on Various Technologies (IPv6 Grid, Secure Communication, etc.)
  - Leader: Shinji SHIMOJO (Osaka U)

  Susume DATE (Osaka U), Toyokazu AKIYAMA (Osaka U), Hiroshi AIKAWA (ITBL), Yutaka UESHIMA (ITBL), Yoshinori SEO (NEC), Akihiro YAMASHITA
- **Computing Grid Technology:** Simultaneous Processing High-Speed Simulation
  - Leader: Haruki NAKAMURA (Osaka U)

  Kizashi YAMAGUCHI (Osaka U), Kenzo AKAZAWA (Osaka U), Yoshihisa KURACHI (Osaka U), Shoji TAKADA (Kobe U), Koichi TSUZUKI (Hitachi), Shirun HO (Hitachi)
- **Data Grid Technology**: Unified Access to Heterogeneous Databases
  - Leader: Hideo MATSUDA (Osaka U)

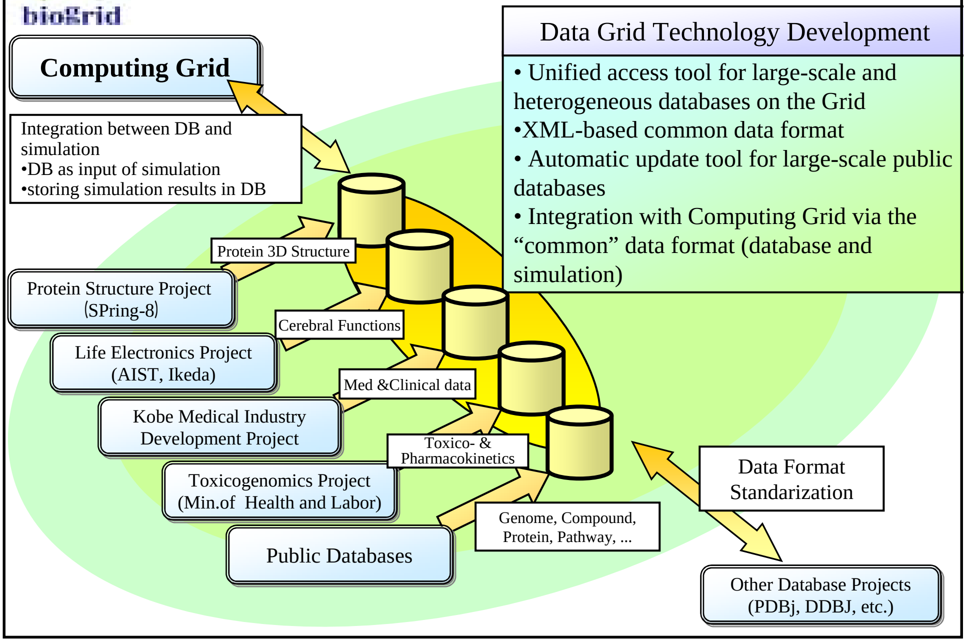
  Teruo YASUNAGA (Osaka U), Takenao OHKAWA (Osaka U), Hiroyuki TOH (Kyoto U), Masaharu ISOYAMA (Protein Research Foundation), Junji Yoshii (Hitachi Software Engineering), Masato KITAJIMA (Fujitsu Kyushu Engineering System), Goichi Nakagawa (Aztec System), Akihiro KOKUFUDA (HP), Yoshinori SATO (Mitsubichi Space Software)
- **Data Online Analysis Technology**: Analyzing Observation Data from Remote Experimental Facilities
  - Leader: Shinji SHIMOJO (Osaka U)

  Naoto YAGI (SPring-8)
- **Business Modeling**: Evaluation of Technology and their Business Transfer
  - Leader: Tsuneaki SAKATA (Osaka U)

  Toshikazu TAKADA (NEC), Yukihiro EGUCHI (Mitsui Knowledge Industry), Yoshitada FUKUOKA (Mitsui Knowledge Industry), Ryuichi MOROTOMI (Life Science Park Center), Toshiyuki OKUMURA (SII), and 7 Pharmaceutical Companies.

Biological Data Grid

Goal of Data Grid (by H. Matsuda)



Computing Grid

Integration between DB and simulation
- DB as input of simulation
- storing simulation results in DB

Data Grid Technology Development

- Unified access tool for large-scale and heterogeneous databases on the Grid
- XML-based common data format
- Automatic update tool for large-scale public databases
- Integration with Computing Grid via the "common" data format (database and simulation)

PDBj

Protein Data Bank Japan

Welcome to the web site of Protein Data Bank Japan (PDBj)

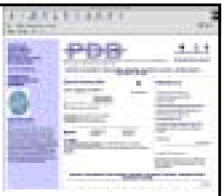
http://www.pdbj.org
and
http://pdbj.protein.osaka-u.ac.jp

PDBj (Protein Data Bank Japan)

http://www.pdbj.org/

Data Curation/Editing/Distribution

(About 10% of the world data are processed for Asia and Oceania.)

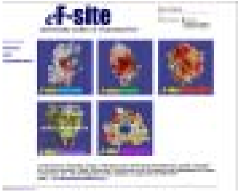


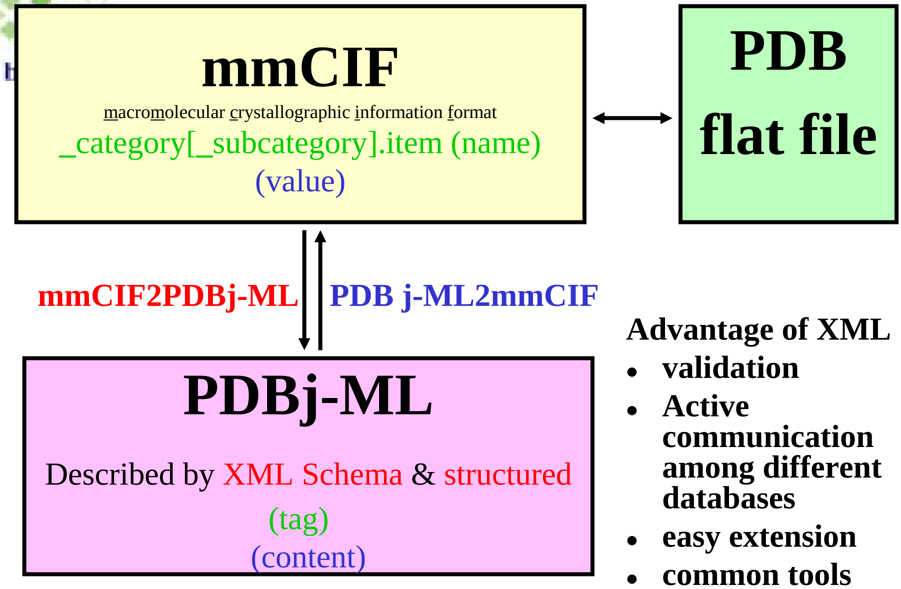
1) Construction of Advanced PDB database

- Development of XML description for PDB data (**PDBj-ML**)
- Addition of Protein Function Information

2) Construction of Secondary Databases

- Protein Molecular Surface Database, **eF-site** (Nakamura & Kinoshita)
- Protein Dynamics Database, **ProMode** (Wako & Endo)





mmCIF

<u>m</u>acro<u>m</u>olecular <u>c</u>rystallographic <u>i</u>nformation <u>f</u>ormat

_category[_subcategory].item (name)

(value)

PDB flat file

mmCIF2PDBj-ML

PDB j-ML2mmCIF

PDBj-ML

Described by XML Schema & structured

(tag)

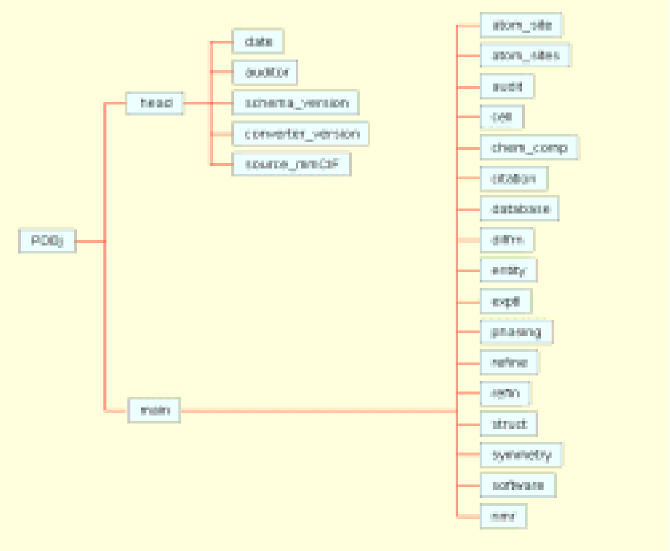
(content)

Advantage of XML

- validation
- Active communication among different databases
- easy extension
- common tools

http://www.pdbj.org/

Structure of PDBj-ML



Example of Secondary Structures

```
HELIX    1   1 ALA A   55  SER A   59  1            6

   <struct>
      <conf id="helix_1"
       conf_type_id="HELX_RH_AL_P">
        <beg_label_comp_id>ALA</beg_label_comp_id>
        <beg_label_asym_id>A</beg_asym_comp_id>
        <beg_label_seq_id>55</beg_label_seq_id>
        <end_label_comp_id>SER</end_label_comp_id>
        <end_label_asym_id>A</end_asym_comp_id>
        <end_label_seq_id>59</end_label_seq_id>
        <details>RIGHT-HANDED ALPHA HELIX</details>
      </conf>
    </struct>
```

Atom Coordinates etc are kept in a separate file

```
<atom_site>
   <ext_file>
     ftp://pdbj.protein.osakau.ac.jp/pdbj_ml/
         all/1crn_atoms.pdbj.gz
   </ext_file>
   <model nid="1" mmCIF_id="1">
      <chain nid="1" mmCIF_id="A">
         <residue nid="1" AorH="ATOM">
            <label_comp_id>THR</label_comp_id>
  :
  :
</atom_site>
```

```
<?xml version="1.0" encoding="UTF-8" ?>
   <PDBj_extension entry="1crn"
       xmlns=http://pdbj.protein.osakau.ac.jp/ns/PDBj-ML-ext
       xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
       xsi:schemaLocation="http://pdbj.protein.osaka-u.ac.jp/ns/PDBj-ML-ext
       http://pdbj.protein.osaka-u.ac.jp/schemas/PDBj-ML-ext/ext_00.xsd">
  <compul_block>

  <!-- [items of atom_record] <atom_record> LabelMdlID LabelAsymID
LabelSeqID LabelAltID LabelCompID TypeSymbol LabelAtomID AuthAtomID CartnX
CartnY CartnZ Occupancy Bfactor</atom_record>   -->

  <atom_record id="1" mmCIF_id="1">1 1 1 ## THR N N N 17.047 14.099 3.625
    1.00 13.79</atom_record>
  <atom_record id="2" mmCIF_id="2">1 1 1 ## THR C CA CA 16.967 12.784 4.338
    1.00 10.80</atom_record>
  <atom_record id="3" mmCIF_id="3">1 1 1 ## THR C C C 15.685 12.755 5.133
    1.00 9.19</atom_record>
  <atom_record id="4" mmCIF_id="4">1 1 1 ## THR O O O 15.268 13.825 5.594
    1.00 9.85</atom_record>
  <atom_record id="5" mmCIF_id="5">1 1 1 ## THR C CB CB 18.170 12.703 5.337
    1.00 13.02</atom_record>
  :
  :
  :
 <compul_block>
```

Comparison of the total file sizes

	Size (GB)	ratio
PDB-flat	10.8	1
mmCIF	13.2	1.22
PDBj-ML	25.7	2.38

Inclusion of functional information

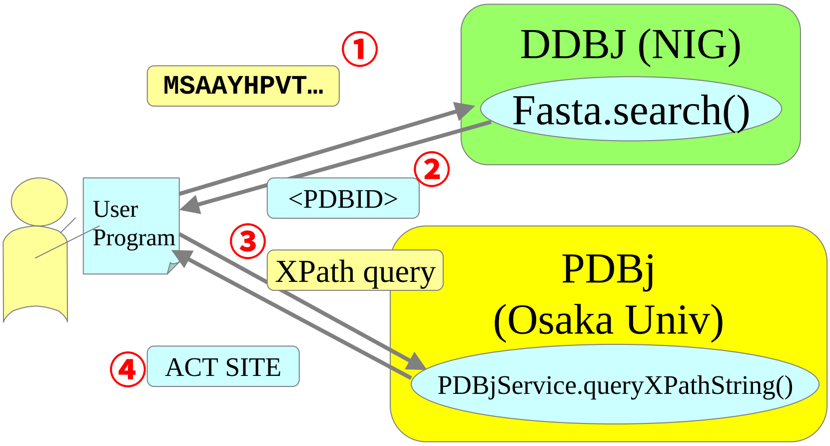
```
<struct>
  <site id="CATRES1" auth_validate="N" info_subtype="catalytic"
      info_type="CATRES" update_id="2">
    <num_residues>3</num_residues>
    <details>a catalytic site defined by CATRESS, Medline 98100076</details>
    <site_gen nid="1">
      <label_comp_id>ARG</label_comp_id>
      <label_asym_id>1</label_asym_id>
      <label_seq_id>100</label_seq_id>
      <details>transition-state stabilisation. stabilise negatively charged
      tetrahedral intermediate</details>
    </site_gen>
    <site_gen nid="2">
      <label_comp_id>ASP</label_comp_id>
      <label_asym_id>1</label_asym_id>
      <label_seq_id>46</label_seq_id>
      <details>acid/base, transition-state stabilisation. stabilises
      positively charged NH3+ part of intermediate, then as a base removes
      proton from this, leading to collapse and formation of asn</details>
    </site_gen>
    <site_gen nid="3">
      <label_comp_id>GLN</label_comp_id>
      <label_asym_id>1</label_asym_id>
      <label_seq_id>116</label_seq_id>
      <details>transition-state stabilisation. stabilise negatively charged
      tetrahedral intermediate</details>
    </site_gen>
  </site>
</struct>
```

Implementation of XPath

Limited XPath search is implemented. For example, the following query will search PDB entries which have more than two types of polypeptides (that is, heterodimers, heterotrimers, complexes etc).

```
 /PDBj
[ count( /PDBj/main/entity/entity_item
  [ formula_weight >= 2000 and poly_type =
    “polypeptide(L)”] ) > 2 ] /@entry_ID
```

WEB service through SOAP



SOAP Query for: Sequence → Structure → Function

SOAP Programing by java (1)

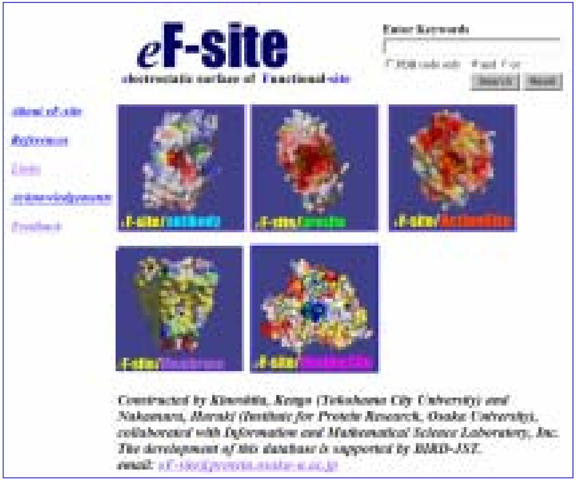
```
public class SoapTest {

  public static void main ( String args[] ) {
         String wsdlURL = "http://xml.nig.ac.jp/wsdl/Fasta.wsdl";
         String nmspace = "http://www.themindelectric.com/wsdl/Fasta/";
         String srvname = "Fasta";
         String fncname = "search";
         String query   = ">Test¥n" + args[0];
         String param   = "-b 1 -d 1 -H";

         QName serviceQN = new QName( nmspace, srvname );
         QName portQN    = new QName( nmspace, srvname );
         Service service = new Service( new URL(wsdlURL), serviceQN );
         Call call = ( Call )service.createCall( portQN, fncname );

         String result = ( String )call.invoke(
                       new Object[] { "fasta3", "PDB", query, param } );

         System.out.println( result );
    }

}
```

SOAP Programing by java (2)

```
public class PDBjXPath {

    public static void main ( String args[] ) {
        try {
            String wsdlURL = "http://www.pdbj.org/wsdl/PDBjSoapService.wsdl";
            String nmspace = "http://www.pdbj.org/soap";
            String srvname = "PDBjSoapService";
            String fncname = "queryXPathString";
            String query   = args[0];
            Integer rstnum  = new Integer( args[1] );

             QName serviceQN = new QName( nmspace, srvname );
            QName portQN    = new QName( nmspace, srvname );
            Service service = new Service( new URL(wsdlURL), serviceQN );

            Call call = ( Call )service.createCall( portQN, fncname );

            String result = (String)call.invoke(
                              new Object[] {query,rstnum} );
            System.out.println( result );
        }
        catch ( Exception e ) {
            e.printStackTrace();
        }
    }
}
```

eF-site *(electrostatic-surface of Functional site)*

http://www.pdbj.org/eF-site/



eF-site

electrostatic surface of Functional-site

Enter Keywords

Constructed by Kinoshita, Kengo (Yokohama City University) and Nakamura, Haruki (Institute for Protein Research, Osaka University), collaborated with Information and Mathematical Science Laboratory, Inc. The development of this database is supported by BIRD-JST.
email:

eF-site
electrostatic-surface of
Functional site

- antibody (103)
- prosite (5411)
- ActiveSite (5047)
- Membrane (51)
- Binding site (1920)

as a total: 7547 entries

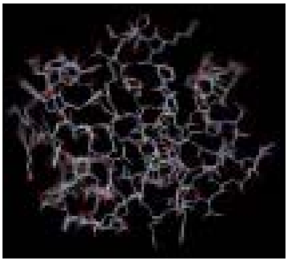
In addition,
eF-site/DNA (101)
eF-site/mono-nucleotide (1044)
are now prepared.

Kinoshita et al., *J. Struct. Funct. Genomics* **2**, 9-22 (2002)

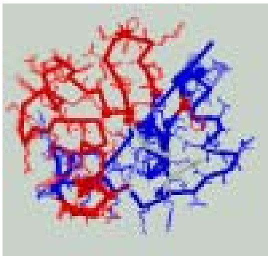
ProMode: Protein Dynamics Database

by Wako & Endo

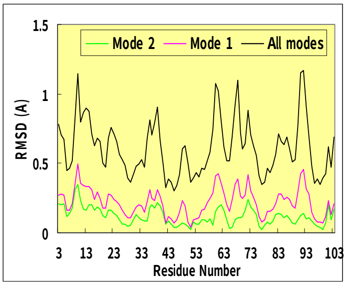
PDB → Regularization Energy minimization → Normal Mode Analysis



Normal mode vibration



Dynamic domains by DynDom

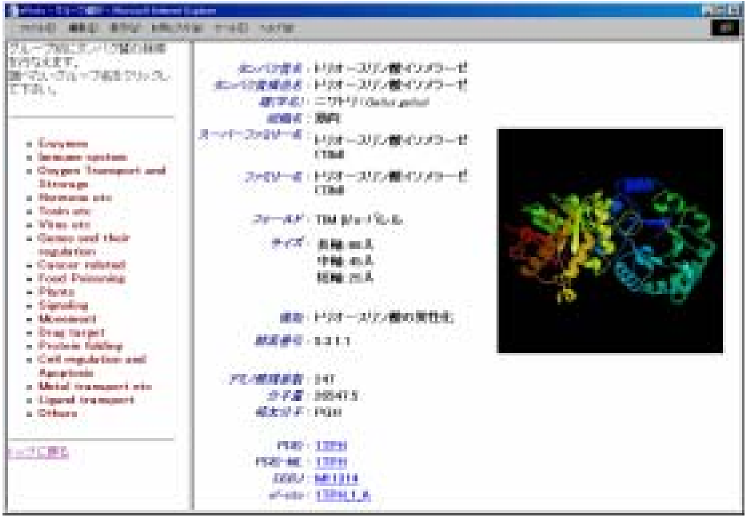


Root-mean-square fluctuations of Cα atoms

eProtS (encyclopedia of Protein Structures)

will appear in this February

English and Japanese versions



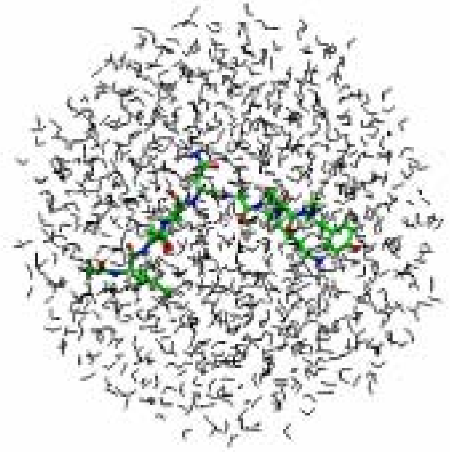
- Enzymes
- Immune system
- Oxygen Transport and Storage
- Hormone etc
- Toxin etc
- Virus etc
- Genes and their regulation
- Cancer related
- Food Poisoning
- Plants
- Signaling
- Movement
- Drug target
- Protein folding
- Cell regulation and Apoptosis
- Metal transport etc
- Lipid transport
- Others

Biological Computing Grid

bioGrid

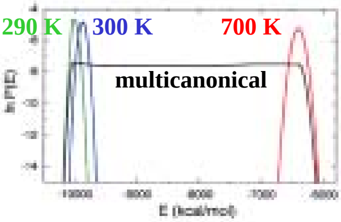
Multicanonical simulation of a β-hairpin peptide in explicit water

Ace-Ile-Thr-Val-Asn-Gly-Lys-Thr-Tyr-NMe



Simulation Conditions:

with 1,060 TIP3P H_2O
Cell Multipole Method
PRESTO ver3 McMD
covers 290-700K
(after 1.2 x 10^8 steps)



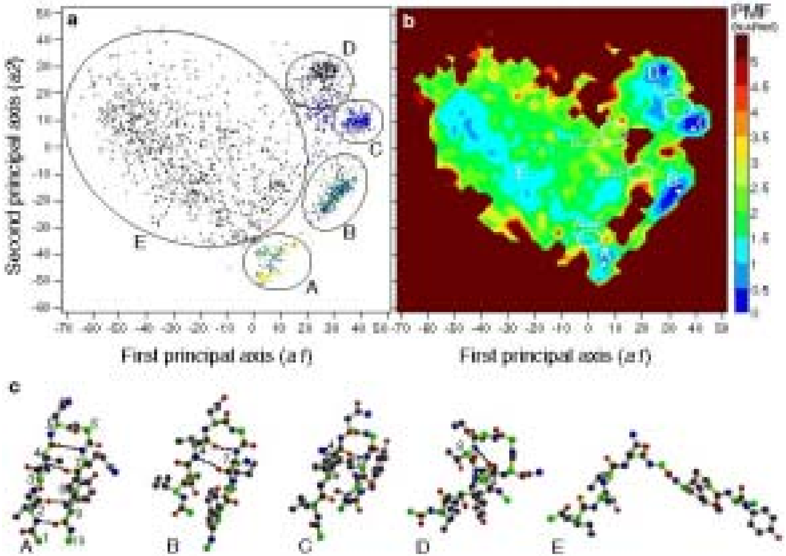
***Higo et al.** (2001) and*
***Kamiya et al.** (2002)*

---

bioGrid

Free energy landscape of a β-hairpin peptide

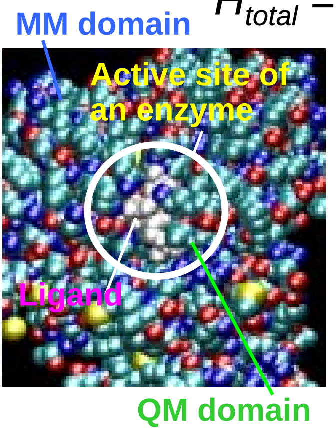
***Kamiya et al.** (2002) Protein Science, 11, 2297-2307.*

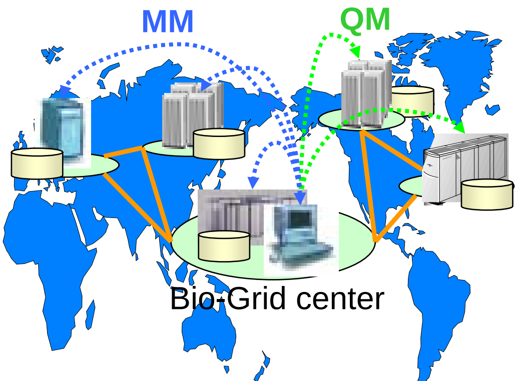


Quantum mechanics (QM) and Molecular mechanics (MM) simulation coupled on Grid environments

Hamiltonian of total system

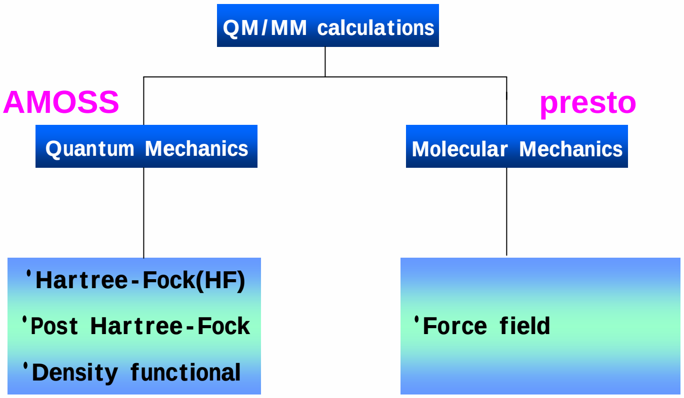
$$H_{total} = H_Q(x,x) + H_{QM}(x,y) + H_M(y,y)$$





Coupled simulation on Grids

Hybrid QM/MM METHOD



QM/MM calculations

AMOSS

Quantum Mechanics

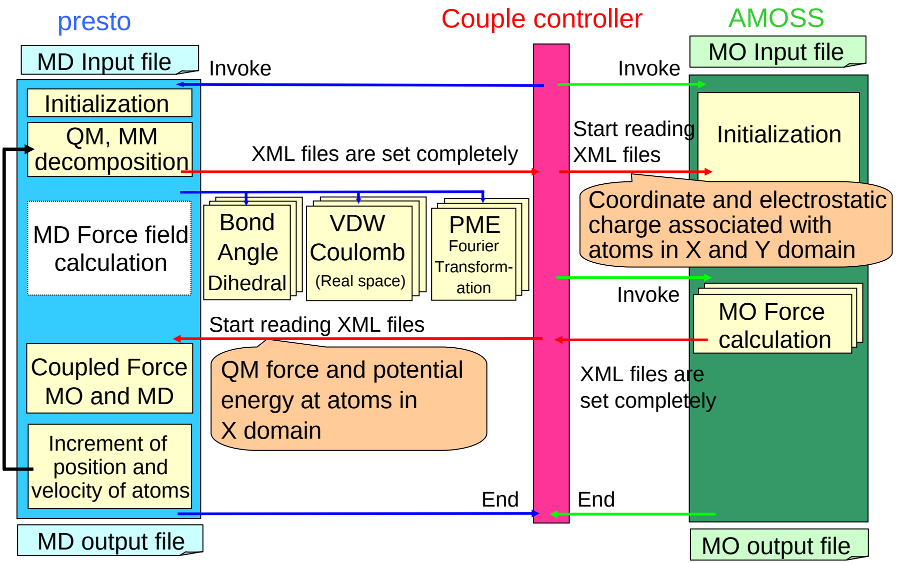
- Hartree-Fock(HF)
- Post Hartree-Fock
- Density functional

presto

Molecular Mechanics

- Force field

Workflow of coupled simulation between presto(MM) and AMOSS(QM)



presto

Couple controller

AMOSS

MD Input file

Invoke

Invoke

MO Input file

Initialization

QM, MM decomposition

XML files are set completely

Start reading XML files

Initialization

Coordinate and electrostatic charge associated with atoms in X and Y domain

MD Force field calculation

Bond Angle Dihedral

VDW Coulomb (Real space)

PME Fourier Transform-ation

Invoke

MO Force calculation

Start reading XML files

QM force and potential energy at atoms in X domain

XML files are set completely

Coupled Force MO and MD

Increment of position and velocity of atoms

End

End

MD output file

MO output file

Execution of Hybrid QM/MM calculation

- Divide QM and MM programs into a set of many pieces

- Data Communication between QM and MM program pieces.

→Standard description for data communication between QM and MM

→Proposal and Design for UDS-XML (Universal Data set-XML) , and Creation of the tools to handle it.

biogrid

UDS-XML (Universal Data Set-XML)

[function name] udsput_xml, udsget_xml

[function] write and parse UDS data in XML.

[form] Text, HexDec, Base64

[attribute] Data unit is always described.

---

biogrid

Example of UDS-XML

UDS-XML (form = Base64) (x4/3)

```
<?xml version=“1.0” encoding=“UTF-8”?>
<uds_data size="1" count="200" unit="AU" form="b64">
   <uds_content> sample_hxd.xml </uds_content>
   <uds_comment> Sample of UDS-XML (hexdec) </uds_comment>
   <uds_array_count>4</uds_array_count>
   <uds_array seq="1" element="character" length="76">
pLWkq6TipMik0qS1pLekzrrupMOkv6XQpaSlyqXqpcehvKW/pPKlqKXzpbOhvKXJp
LekxqS9pM6l</uds_array>
   <uds_array seq="2" element="character" length="76">
x6G8pb+k8qPYo82jzLfBvLCkx6XVpaGlpKXrpMvK3cK4pLmk66XXpe2lsKXppeCkz
qXGpbmlyKTH</uds_array>
   <uds_array seq="3" element="character" length="76">p
Lmho6XXpe2lsKXppeCkrMC1vu+ky8awpKSkxqSkpOu+7LnnpM+ks6TOyrikrMC1p
Lekr8bJpOGk</uds_array>
   <uds_array seq="4" element="character" length="40">xqSkpOukz6S6pMekuaG
jpMmkpqTHpLek56Smoak=</uds_array>
</uds_data>
```

Acknowledgements

Nobutoshi Ito, Hisashi Sakamoto, Kaori Kobayashi *(JST)*

Nobuyuki Nakajima *(IPR, Osaka Univ.)*

Junichi Higo *(Tokyp Univ. Pharma.)*

Narutoshi Kamiya *(BERI)*

Yoshifumi Fukunishi *(BIRC-AIST)*

Takashi Kurosawa, Shurin Ho *(Hitachi)*

Toshihiro Sakuma, Toshikazau Takada *(NEC)*

Kizashi Yamaguchi, Kenzo Akazawa, Yoshihisa Kurachi,

Hideo Matsuda, Shinji Shimojyo, Tsuneaki Sakata *(Osaka Univ.)*

Shoji Takada *(Kobe Univ.)*