Protein Informatics: From Sequence to Structure and Function in Proteomics Research

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/
http://www.pdbj.org/
Modeling techniques:
1) motivation
2) method
3) application
4) future
Roles of *Protein Informatics*

- **Construction of Database:**
  - Systematic Analysis & Applications

- **Prediction:**
  - Sequence $\rightarrow$ Structure $\rightarrow$ Function
    - (Genotype) $\rightarrow$ (Proteome) $\rightarrow$ (Phenotype)

- **Simulation:**
  - Molecular simulation
  - Network simulation
  - Simulation for higher levels of system, organelle, cell, and body.
Structural Database
New International organization is just founded in Autumn, 2003.
1) Construction of Advanced PDB database
   • Development of XML description for PDB data \( (\text{PDB-XML}) \)
   • Addition of Exp. Condition and Biochem. Function Information

2) Construction of Secondary Databases
   • Protein Molecular Surface Database, \( e\text{F-site} \) (Nakamura & Kinoshita)
   • Protein Dynamics Database, \( \text{ProMode} \) (Wako & Endo)
   • Encyclopedia of Protein Structures, \( \text{eProtS} \) (Nakamura & Ito)
Structure: MYOGLOBIN
Size [Å]:
- Long Axis: 36
- Middle Axis: 33
- Short Axis: 20
Function: OXYGEN STORAGE
Number of amino acid residues: 153
Molecular weight [Da]: 17855.7

PDB: 1MBN
PDBeML: 1MBN

Display 3D Structure
PDBViewer is used to display 3D structures. Java(TM) Plug-in 1.4 and Java2D 1.3 must be installed on your computer.

to some proteins through their classes.
Click the class of your interest.

- Enzymes
- Immune system
- Oxygen Transport and Storage
  - Deoxyhemoglobin S
  - Myoglobin
- 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
- Ligand transport
- Others

Back to the top page
PDBjViewer

- Offers interactive molecular visualization with the RasMol type commands.
- Can be used both as Stand-alone and as applet with Java3D.
- Any polygons defined by XML can be displayed and manipulated quickly.
- Can parse a PDB-XML file and display the molecule.

http://www.pdbj.org/PDBjViewer/
Structural Bioinformatics: Database Analysis
Structural Genomics/Structural Proteomics: Structural Biology for Protein Networks

Purpose: To Understand Genome Information through Protein 3D Structure

Protein Network analysis

Structural Genomics Project
Protein Sequence → Protein 3D structure → Biological Function

Protein Network
Protein Biochemical Function → Protein Biological Function

N-acetylgalactosamine-6S sulfatase (GALNS)

(Sukegawa et al. (2000) Human Mol. Genetics, 9, 1283-1290)

Protein 3D structures in PDBj → Homology Model

Identify protein function from

- Sequence similarity
- Fold similarity
- Surface similarity

1. Construction of a database for protein surfaces around the functional sites, ‘eF-site (electrostatic-surface of Functional site)’

2. Identification of the similar surface geometry and physico-chemical property by Clique detection algorithm
**eF-site database**

http://www.pdbj.org/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.

1. Construction of a database for protein surfaces around the functional sites, ‘eF-site (electrostatic-surface of Functional site)’

2. Identification of the similar surface geometry and physico-chemical property by Clique detection algorithm  
Clique detection algorithm

(Kinoshita et al., 2002)

**Node** is created for a vertex pair having similar properties & curvatures

**Edge** is created for a pair of nodes having the similar spatial arrangement of the normal vector pairs.

**Clique** is a subgraph of an undirected graph where every node is connected.

→ Search the largest clique (Bron & Kerbosch, 1973)
Structure-based identification of a novel NTPase from *Methanococcus jannaschii*

Kwang Yeon Hwang\(^1,2\), Ji Hyung Chung\(^1-3\), Sung-Hou Kim\(^4\), Ye Sun Han\(^2\) and Yunje Cho\(^2\)

Almost half of the entire set of predicted genomic products from *Methanococcus jannaschii* are classified as functionally unknown hypothetical proteins. We present a structure-based identification of the biochemical function of a protein with an as yet unknown function from a *M. jannaschii* gene, Mj0226. The crystal structure of Mj0226 protein determined at 2.2 Å resolution reveals that the protein is a homodimer and each monomer folds into an elongated α/β structure of a new fold family. Comparisons of Mj0226 protein with protein structures in the database, however, indicate that one part of the protein is homologous to some of the nucleotide-binding proteins. Biochemical analysis shows that Mj0226 protein is a novel nucleotide triphosphatase that can efficiently hydrolyze nonstandard nucleotides such as XTP to XMP or ITP to IMP, but not the standard nucleotides, in the presence of Mg\(^{2+}\) or Mn\(^{2+}\) ions.
Application to a hypothetical protein: MJ0226


Search result against eF-site/ActiveSite + eF-site/mononucleotide (1684 entries)

Query: MJ0226 free form

Similarity was found here.

Folylpolyglutamate synthetase

Pyruvate kinase
A function unknown protein, TT1542, from *Termus thermopilus* HB8

Its structure is determined by N. Handa & S. Yokomaya (RIKEN)

- 227 amino-acid residues
- There are more than 100 proteins found in prokaryotes and eukaryotes, whose functions are unknown.
- The fold is Rossmann fold-like topology.

Similarity search against 22747 binding sites in eF-site.
13 putative binding site

- The 31 ‘significantly’ similar binding sites are classified into 13 binding sites according to their position on the query protein.

- 10th binding pocket is the most promising.
  - from the highest surface similarity.
  - from existence of a cluster of conserved residues.

Conserved sequence

From reverse angle
10\textsuperscript{th} putative binding site

4 kinds of proteins are found to be similar.

1. a.43.1.2: 1mjq, 1mjl, 1mjq with SAM
2. d.92.1.11: 1b3d, 1d5j, 1d7x, 966c with MM3, SPC, RS2, S27
3. b.47.1.2: 1f0r with 815
4. b.71.1.1: 1jdd with GLC

All of the putative ligand are kinds of sugar related compound.
Binding mode found in 10\textsuperscript{th} binding site
Ligand-Protein interaction is, in general, not a rigid body association.
Structural Bioinformatics: Molecular Simulations
Analysis of Free Energy Landscape

(Free energy distribution in the conformational space)

Search Conformations as exhaustively as possible, and reproduce Boltzmann-Gibbs ensemble near 300 K by enhanced conformational sampling

Multicanonical MD (Nakajima et al., 1997)

Tsallis Dynamics method (Fukuda & Nakamura, 2002)
Multicanonical simulation of a $\beta$-hairpin peptide in explicit water

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

Simulation Conditions:
with 1,060 TIP3P H$_2$O
AMBER parm96
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)
Free energy landscape of a β-hairpin peptide
Force-biased iteration scheme for McMD: FB-McMD

- Random walk in energy space
- Automatic determination of force-scaling factor

\[ \dot{\rho}_a = \nu(E) \vec{F}_a \quad (\nu(E) \neq 1) \]

\[ \nu^{i+1}(E) = \nu^i(E) + \Gamma^i(E) \]

\[ \Gamma^i(E) = k_B T_0 \left[ \frac{\partial \ln P^i(E)}{\partial E} \right] \]

Applications to:

- Prediction of Peptide Conformations,
- Dynamic Modeling of Flexible Loops,

and

- Ligand/Inhibitor Docking

→ *in silico* drug screening
Collaborators

Yuko Tuchiya, Nobuyuki Nakajima
(Inst. Protein Res., Osaka Univ.)

Nobutoshi Ito
(PDBj, Tokyo Univ. Med. Dent.)

Kaori Kobayashi
(PDBj, BIRD-JST)

Kengo Kinoshita
(Yokohama City Univ.)

Jae Gil Kim, Yukihisa Watanabe,
Yoshifumi Fukunishi
(JBIRC-AIST)