*J. Biochem.* **135**, 79–84 (2004)
DOI: 10.1093/jb/mvh009

# Het-PDB Navi.: A Database for Protein–Small Molecule Interactions

**Akihiro Yamaguchi[1], Kei Iida[*,2], Nobuaki Matsui[2], Shirou Tomoda[3], Kei Yura[4] and Mitiko Go[†,1]**

*[1]Department of Bio-Science, Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology, 1266, Tamura-cho, Nagahama, Shiga, 526-0829; [2]Division of Biological Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8602; [3]Department of Electronic and Information System Engineering, Faculty of Science and Technology, Hirosaki University, Hirosaki, 036-8560; [4]Quantum Bioinformatics Group, Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, 8-1 Umemidai, Kizu, Souraku, Kyoto, 619-0215*

Received June 22, 2003; accepted November 13, 2003

**The genomes of more than 100 species have been sequenced, and the biological functions of encoded proteins are now actively being researched. Protein function is based on interactions between proteins and other molecules. One approach to assuming protein function based on genomic sequence is to predict interactions between an encoded protein and other molecules. As a data source for such predictions, knowledge regarding known protein-small molecule interactions needs to be compiled. We have, therefore, surveyed interactions between proteins and other molecules in Protein Data Bank (PDB), the protein three-dimensional (3D) structure database. Among 20,685 entries in PDB (April, 2003), 4,189 types of small molecules were found to interact with proteins. Biologically relevant small molecules most often found in PDB were metal ions, such as calcium, zinc, and magnesium. Sugars and nucleotides were the next most common. These molecules are known to act as cofactors for enzymes and/or stabilizers of proteins. In each case of interactions between a protein and small molecule, we found preferred amino acid residues at the interaction sites. These preferences can be the basis for predicting protein function from genomic sequence and protein 3D structures. The data pertaining to these small molecules were collected in a database named Het-PDB Navi., which is freely available at http://daisy.nagahama-i-bio.ac.jp/golab/hetpdbnavi.html and linked to the official PDB home page.**

**Key words: co-factor, database, metal, nucleotide, protein 3D-structure, protein function, sugar.**

Abbreviations: 3D, three-dimensional; PLP, pyridoxal phosphate; PDB, Protein Data Bank.

One of the challenges in bioinformatics is to assign functions to proteins encoded in predicted open reading frames in genomic sequences. The current progress in genome sequencing projects has provided more than 20 billion nucleotides of DNA sequence (*1*). An enormous number of proteins are encoded by these DNA sequences, and computational methods have been developed to obtain hints about their functions (*2*). Even with such efforts, the functions of about half of the proteins predicted from genomic sequences remain unknown (*3*).

Protein function is based on interactions between proteins and other molecules. Enzymes interact with cofactors and substrates to catalyze chemical reactions, signal transducers interact with other proteins, and transcription factors interact with DNA to regulate transcription. In order to study the biological function of a predicted protein, knowledge of the cofactors that make the protein fold and work is required. An understanding of a protein's modifications and interaction partners is also necessary. In short, one way to assume the biological function of proteins is to predict the interactions between proteins and other molecules. A number of methods for predicting these interactions have been developed (*4*), but effective methods for postulating biological function remain to be established.

Fundamental information regarding interactions between proteins and other molecules is found in PDB (*5*). Protein 3D structures are often determined in complexes with other molecules called heterogen molecules. The same heterogen molecules are often found complexed with different proteins. Therefore, atomic interactions between small molecules and proteins can be quantitatively assessed, and a general tendency in interactions can be deduced. Classification of protein-heterogen molecule interactions is an initial step in understanding preferences in interactions and for developing novel methods for predicting protein function based on interactions with other molecules.

We have therefore, developed a heterogen molecule database, named Het-PDB Navi., to survey heterogen molecules in PDB and identify molecules for which

[*]Present address: Genomic Knowledge Base Research Team, Bioinformatics Group, Genomic Sciences Center, RIKEN, 1-7-22, Suehiro, Tsurumi, Yokohama, Kanagawa, 230-0045
[†]To whom correspondence should be addressed. Tel: +81-749-64-8127, Fax: +81-749-64-8126, E-mail m_goh@nagahama-i-bio.ac.jp

 

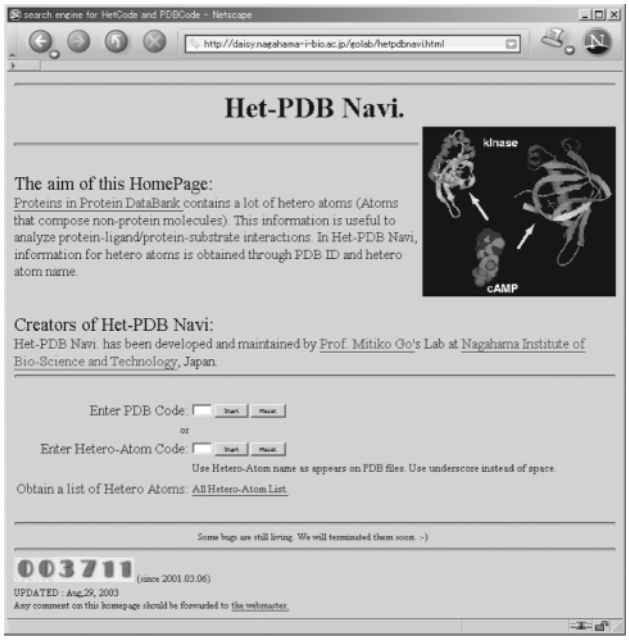


Fig. 1. **The top page of Het-PDB Navi. available at http://daisy.nagahama-i-bio.ac.jp/golab/hetpdbnavi.html.**

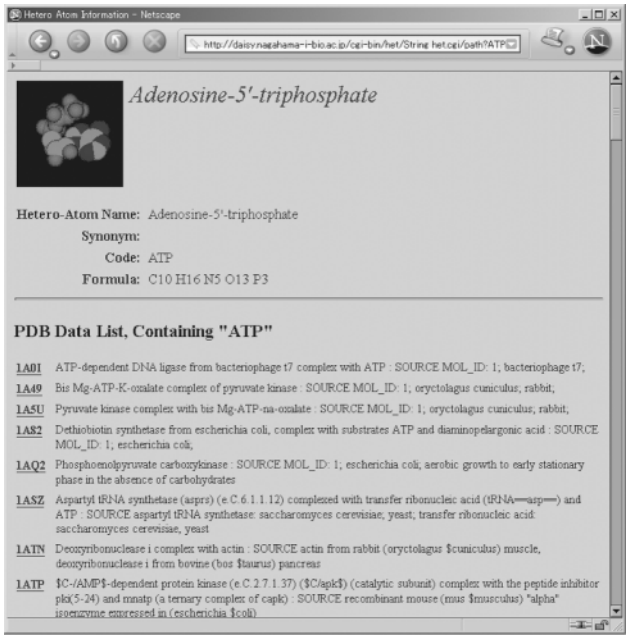


Fig. 2. **Result of a heterogen molecule search.** When a user searches PDB entries containing an ATP molecule, Het-PDB Navi. displays all the entries containing ATP with a link to the 3D structure of each entry.

enough data is available to study their interactions statistically. With Het-PDB Navi., one can survey PDB using the name of a heterogen molecule. The information given in Het-PDB Navi. is similar to that in PDBsum (*6*), but Het-PDB Navi. also provides lists of heterogen molecules grouped by their names. This article further deals with the preferences of amino acid residues at interaction sites for heterogen molecules.

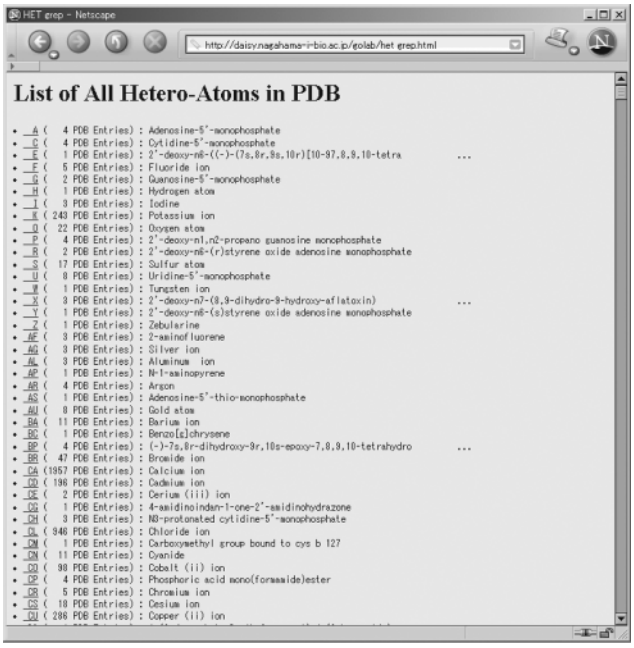


Fig. 3. **A list of heterogen molecules in PDB.** The list contains three-letter identification of heterogen molecules, the number of PDB entries containing heterogen molecules, and the description of the molecules. The three-letter identification is linked to a list of PDB entries containing the heterogen molecule (Fig. 2).

## MATERIALS AND METHODS

*Extraction of Names of Small Molecules from PDB*—Names and three-letter identification of heterogen molecules that interact with proteins were extracted from PDB (*5*). Each entry in PDB contains line-by-line information, and lines with HETNAM or HETATM in the first column contain three-letter identification and names of heterogen molecules. Heterogen molecules were classified based on three-letter identification and stored in the database. Chemical modifications of proteins, such as phosphorylation and sugar attachment, were treated as heterogen molecules. Water molecules were not considered as such.

*Architecture of Het-PDB Navi.*—Het-PDB Navi. is a text-based perl-driven database. The present Het-PDB Navi. accepts queries about heterogen molecules by three-letter identification or PDB ID, and then searches PDB entries (Fig. 1). When three-letter identification of a heterogen molecule is used as a query, all PDB entries with that heterogen molecule are shown (Fig. 2). The 3D structures of each entry can be seen by clicking the PDB ID, when Rasmol (*7*) is installed as a plug-in of the browser. When a protein is an enzyme, the database also shows the chemical reaction that the protein catalyzes. With this data and list of interacting molecules, it is possible to assume whether the molecule is a cofactor or a substrate. A link to PDBsum (*6*) is also found here.

The database contains a list of three-letter identifications of heterogen molecules with the name of each molecule (Fig. 3). When the three-letter identification of a heterogen molecule is unknown, it can be found in the list given in the database. The database contains all PDB

entries, and therefore, there are cases when identical protein-small molecule interactions are included multiple times. We considered that even with the same protein and heterogen molecule pair, the interactions between them may change due to crystallization conditions, and these differences may be biologically important. There are other cases where identical proteins have different heterogen molecules. Therefore, we did not select proteins by sequence identity. The database is freely available under the URL: http://daisy.nagahama-i-bio.ac.jp/golab/hetpdbnavi.html, and linked to the official PDB home page. Data in Het-PDB Navi. are updated regularly.

*Statistics of Heterogen Atoms in PDB Based on Het-PDB Navi.*—The same heterogen molecules in PDB are often annotated with different three-letter identifications. Therefore, classification of heterogen molecules was first done by automatically finding identical matches of three-letter identifications and then by manually comparing real names of heterogen molecules. Statistical analyses on preferred amino acid residues and atoms that interact with heterogen molecules were carried out for molecules with a sufficiently large number of interactions. Protein-heterogen molecule interactions were defined as follows. A protein atom was determined to be bound to a metal ion, when the distance between them was less than the sum of the van der Waal radii of both atoms. For other heterogen molecules, a protein atom residing less than 4.0Å away from one of the heterogen molecule atoms was defined to interact with the molecule.

For statistical analyses, identical interactions between identical proteins should not be counted twice. We employed the following steps to eliminate duplicate entries: (i) retrieve all PDB entries that interact with certain heterogen atoms from Het-PDB Navi., (ii) calculate all-against-all identity of the amino acid sequences given in each entry and form clusters by a 30% sequence identity threshold, (iii) build a multiple alignment in each cluster, (iv) eliminate duplicate entries with identical modes of binding to heterogen atoms with identical amino acid residues, and (v) count heterogen atoms and residues interacting with the heterogen atom found in entries remaining in the multiple alignment.

## RESULTS AND DISCUSSION

*Statistics in Het-PDB Navi.*—In the April 2003 release of PDB, there were 35,557 heterogen molecules. These heterogen molecules were placed into 4,189 groups based on three-letter identification and chemical similarity. In PDB, similar molecules, such as hemes, are sometimes annotated by different three-letter identifications, like HEM, HEG, and HEC. Twenty-nine heterogens were found to account for about 64% of the heterogen molecules in PDB (Table 1). In table 1, numbers in parentheses in the second column were obtained by eliminating identical entries. There were no obvious differences in fractions of heterogen atoms. Table 1 contains sulfate, chloride, sodium, glycerol, acetate, and phosphate ions, as well as ethanol and pentanediol, which are presumably agents for protein crystallization. Many of these molecules were artificially added and therefore not considered to possess biological function. However, a case has been reported in which a sulfate ion was located at the surface of the protein where phosphorous atoms from a DNA molecule were located in biologically active form (*8*). Sequence and structural motifs for nucleotide-binding sites were often found around a phosphate group, as in the case of P-loops (*9*), phosphate-binding helix-turn-helix modules (*10*), and others. Therefore, binding sites for artificial substances may also be biologically significant sites.

Table 1. **The 29 most common heterogen molecules.**

| Molecule | Number | Fraction (%) | Class |
|---|---|---|---|
| Sulfate | 6,133 (4,969) | 7.6 | inorganic |
| Mg | 6,010 (1,868) | 7.4 | metal |
| Ca | 4,898 (1,972) | 6.1 | metal |
| Glucosamine | 4,434 (2,814) | 5.5 | sugar |
| Zn | 3,600 (1,608) | 4.5 | metal |
| Porphyrin | 3,314 (2,565) | 4.1 | heme |
| Cl | 2,361 (1,286) | 2.9 | non-metal ion |
| Na | 2,133 ( 679) | 2.6 | metal |
| Mannose | 1,859 ( 627) | 2.3 | sugar |
| Glycerol | 1,787 (1,763) | 2.2 | inorganic |
| Mn | 1,731 ( 536) | 2.1 | metal |
| Fe | 1,612 ( 716) | 2.0 | metal |
| Acetate | 1,403 ( 587) | 1.7 | inorganic |
| Glucose | 1,346 ( 865) | 1.7 | sugar |
| Phosphate | 1,231 ( 874) | 1.5 | inorganic |
| ATP/ADP/AMP | 1,102 ( 792) | 1.4 | nucleotide |
| $NAD^+$/NADPH | 1,098 ( 848) | 1.4 | nucleotide |
| Cd | 903 ( 614) | 1.1 | metal |
| Cu | 881 ( 235) | 1.1 | metal |
| K | 842 ( 749) | 1.0 | metal |
| FAD/FMN | 810 ( 561) | 1.0 | nucleotide |
| GTP/GDP/GMP | 652 ( 291) | 0.8 | nucleotide |
| Galactose | 615 ( 328) | 0.8 | sugar |
| Ethanool | 587 ( 232) | 0.7 | inorganic |
| PLP | 561 ( 197) | 0.7 | nucleotide |
| Hg | 514 ( 399) | 0.6 | metal |
| Pentanediol | 402 ( 402) | 0.5 | inorganic |
| Co | 344 ( 129) | 0.4 | metal |
| Fucose | 325 ( 149) | 0.4 | sugar |
| | 53,488 | 64.3 | |

Table 1 shows the abundance of ions followed by nucleotide-related and sugar-related molecules. The abundance of these molecules allowed us to perform statistical analyses on the interactions between these molecules and proteins.

*Preferences for Metal-Ligating Residues*—Of the top 29 heterogen molecules in Table 1, 11 were metal ions. These metal ions are important components of enzyme active sites or stabilizers of protein 3D structures (*11*). When metal-ligating residues were analyzed for non-homologous proteins in PDB, every metal showed a different preference (Fig. 4).

In PDB, there were 6,010 magnesium ions. There were 1,868 non-homologous magnesium-binding sites, and the metal was bound to 8,984 atoms. Of these, more than half were an oxygen atom of water. The next most common group of atoms that bound a magnesium ion belonged to nucleotides such as ATP and GTP. Ligation by nucleotides was characteristic of this ion. The major atom that bound the ion was an oxygen atom in an aspartic acid

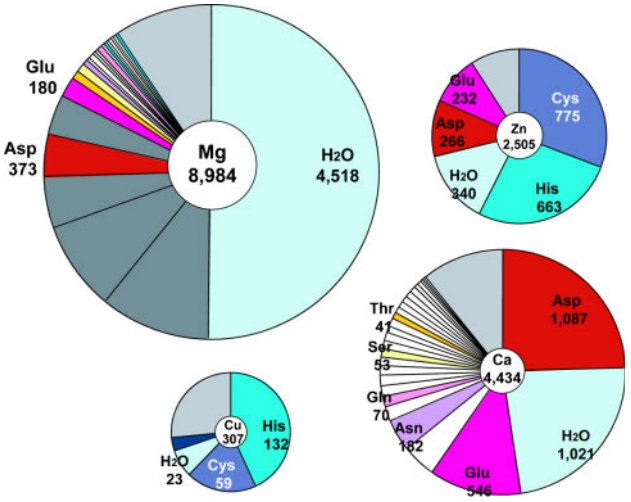


Fig. 4. **Preference of four metals for residues in metal-binding sites of proteins.** Numbers in each pie chart indicate the number of interactions. Different colors represent different types of residues: dark gray, nucleotide-related molecules such as ATP/GTP; white, protein main chain atoms; and light gray (at the top left), the sum of minor residues.

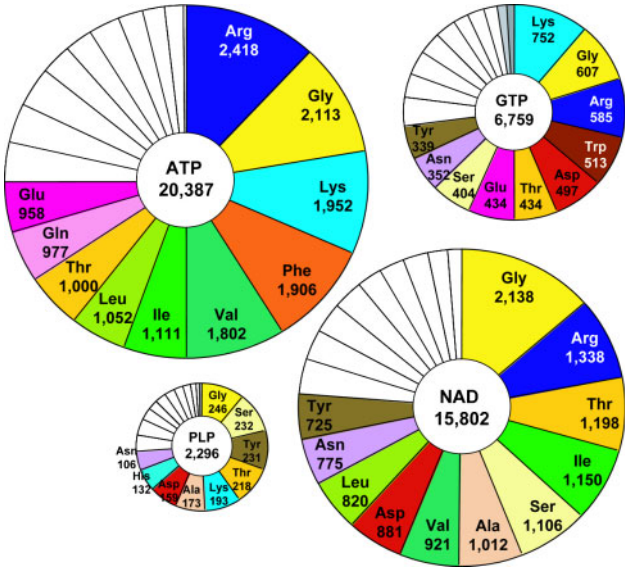


**Fig. 5. Preference for residues in ATP/GTP/NAD+/PLP-binding sites of proteins.** Numbers in each pie chart indicate the number of interactions. Different colors represent different types of residues. The white segment at the top left represents minor residues.

residue side chain. Aspartic acid is considered to be a dominant residue that ligates magnesium ions, and the metal has been found in the active site of DNA polymerase (*12*). The oxygen atom in the glutamic acid side chain was another major ligand for magnesium ions. The number of glutamic acid residues was about half the number of aspartic acid residues (Fig. 4). These statistics will be important for predicting magnesium-binding sites from genomic sequence and 3D protein structure (*11*, *13*).

The second most common metal ion found in PDB was calcium. Of 4,898 calcium ions, 1,972 were bound to non-homologous proteins with 4,434 ligating atoms. The major ligand atoms belonged to side chains of aspartic or glutamic acid residues and water molecules (Fig. 4). A unique aspect of the calcium ion ligand was found in the dominance of amino acid main chain carbonyl groups (*14*). About one quarter of calcium ligands were oxygen atoms from carbonyl groups in the main chain of the protein. The reason for this dominance has not been explained. We checked for relationships between ligation by main chain oxygen and protein fold or function, but found no correlations. Frequent ligation by main chain oxygen atoms of calcium ion ligands is one of the causes of difficulty in predicting calcium-binding sites from amino acid sequences.

The third most common group was a zinc ion, of which 3,600 existed, which were grouped into 1,608 zinc ion binding sites with 2,505 ligating atoms. Zinc is considered to be a soft ion, similar to a copper ion, and different from magnesium and calcium ions, which are called hard ions (*15*). The difference in the chemical characteristics of ions is reflected by differences in ligating atoms. The majority of atoms that bound zinc ions were sulfur atoms in cysteine side chains and nitrogen atoms in histidine side chains (Fig. 4). A similar tendency is found for copper ions. However, one quarter of atoms complexed with zinc ions were oxygen atoms in asparatic acid, glutamic acid, or water. Zinc ions have been considered to be rarely ligated by oxygen atoms, but our survey results challenge this concept. In the case of copper ions, ligation by oxygen is rarely observed. The contribution from aspartic and glutamic acid residues to zinc ligation was similar to that for magnesium and calcium ions, but these two ions are rarely ligated by nitrogen atoms.

*Preferences for Nucleotide-Binding Residues*—The second largest group in Table 1 was nucleotides. Studies on binding sites for ATP/GTP have been reported (*16*). The residues that interact with ATP and GTP have been reported to be similar, despite the fact that ATP and GTP mediate distinct biological functions (*16*). In our study, the three most common residues that interact with ATP and GTP were the same as those from a previous study, namely, arginine, lysine, and glycine (Fig. 5). The first two are positively charged and are expected to interact with phosphorus groups of ATP and GTP. Glycine residues have often been found at ATP- and GTP-binding sites, which are supposed to interact with phosphorus

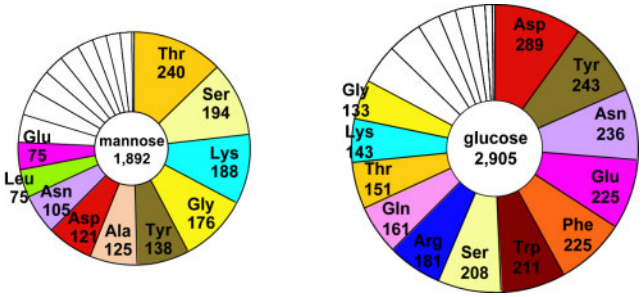


Fig. 6. **Preference of mannose and glucose for residues in sugar-binding sites of proteins.** Numbers in each pie chart indicate the number of interactions. Different colors represent different types of residues. The white segment at the top left represents minor residuess.

groups. Following the top three residues, however, the next most common residues differed for ATP- and GTP-binding sites. Phenylalanine and aliphatic residues were often found at ATP-binding sites, whereas tryptophan, aspartic acid, and glutamic acids were most often found at GTP-binding sites (Fig. 5). These differences, reported here for the first time, will contribute to predicting ATP- and GTP-binding sites in genome sequences and 3D structure of proteins.

Binding sites for $NAD^+$ and NADPH molecules were often found in PDB and have been expected to have similar residue preferences to those for ATP, because both molecules contain an adenine ring with a phosphorous group. As expected, both binding sites were abundant in glycine and arginine residues (Fig. 5). However, an abundance of lysine and phenylalanine, which were evident in ATP-binding sites, was not observed in $NAD^+$-binding sites.

Pyridoxal phosphate (PLP) is another nucleotide-like molecule often found in PDB (Table 1). At the PLP-binding sites, glycine, serine, and tyrosine residues were often found. An abundance of glycine residue was also found in ATP, GTP, and $NAD^+$-binding sites, but an abundance of serine and tyrosine residues was specific to PLP-binding sites. The large number of lysine residues was explained by the fact that PLP often formed covalent interactions with protein molecules at lysine residue side chains (*17*). PLP is a molecule with phosphorus groups and hence expected to have electrostatic interactions with arginine residues. However, binding sites for PLP were not rich in arginine.

*Preferences for Sugar-Binding Residues*—Table 1 contained a reasonable number of sugar molecules, namely, glucosamine, mannose, glucose, galactose, and fucose. These sugar molecules are either enzyme substrates or products or sugars that are attached to protein surfaces. Interactions between sugar molecules, such as mannose or glucose, and proteins are known to be important for cell-cell interactions (*18*). However, few statistical analyses of sugar-protein interactions have been performed thus far. Figure 6 illustrates differences in amino acid residue preferences at mannose- and glucose-binding sites. In PDB, there were 1,859 mannose molecules (Table 1). Homologous interactions between mannose and proteins were eliminated by sequence identity between proteins, resulting in a total of 1,892 interacting residues. Mannose is known to bind covalently to an aspartic acid residue side chain. However, we found no obvious preferential interactions between mannose and aspartic acid residues. The interactions showed no obvious preferences in amino acid residues.

Glucose is a two-epimer of mannose with exactly the same molecular weight as mannose. This chemical property suggested that both sugars might interact with proteins in a similar way. In PDB, there were 1,346 glucose molecules (Table 1) and 2,905 interacting residues. In the case of glucose-protein interactions, residues with negative charges or aromatic rings were preferred (Fig. 6). Differences in interactions between sugars and proteins may relate to functional differences in the proteins that interact with the sugar. For instance, mannose is often found in glycoproteins, while glucose takes part in cellular respiration and metabolic pathways. Knowledge of the preferred residues for interactions between proteins and specific sugar molecules will pave the way for predicting and designing protein-sugar interactions.

Genome sequencing and structural genomics projects will supply a tremendous amount of data on the biological functions of proteins. In PDB, we now have enough structural data to study interactions between proteins and metals, nucleotides, or sugars. The discovery of patterns in these interactions may provide a basis for predicting interactions between small molecules and proteins. Het-PDB Navi. will provide basic knowledge on interactions between proteins and small molecules.

This work was supported by a Grant-in-Aid for Scientific Research on Priority Area (C) Genome Information Science from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan and by a short-term grant for Research and Development for Applying Advanced Computational Science and Technology of the Japan Science and Technology (JST) Corporation.

REFERENCES

1. Nierman, W.C., Eisen, J.A., Fleischmann, R.D., and Fraser, C.M. (2000) Genome data: what do we learn? *Curr. Opin. Struct. Biol.* **10**, 343–348
2. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707–725
3. Kim, S.-H. (2000) Structural genomics of microbes: an objective. *Curr. Opin. Struct. Biol.* **10**, 380–383
4. Aloy, P. and Russell, R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA* **99**, 5896–5901
5. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242
6. Laskowski, R.A, Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L., and Thornton, J.M. (1997) PDBsum: A Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **22**, 488–490
7. Bernstein, H.J. (2000) Recent changes to RasMol, recombining the variants. *Trends Biochem. Sci.* **25**, 453–455
8. Nakagawa, N., Sugahara, M., Masui, R., Kato, R., Fukuyama, K., and Kuramitsu, S. (1999) Crystal structure of Thermus thermophilus HB8 UvrB protein, a key enzyme of nucleotide excision repair. *J. Biochem.* **126**, 986–990
9. Walker, J.E., Saraste, M., Runswick, M.J., and Gay, N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951
10. Yura, K., Shionyu, M., Kawatani, K., and Go, M. (1999) Repetitive use of a phosphate-binding module in DNA polymerase β, Oct-1 POU domain and phage repressors. *Cell. Mol. Life Sci.* **55**, 472–486
11. Bertini, I. and Rosato, A. (2003) Bioinorganic chemistry in the postgenomic era. *Proc. Natl Acad. Sci. USA* **100**, 3601–3604
12. Joyce, C.M. and Steitz, T.A. (1994) Function and structure relationships in DNA polymerases. *Annu. Rev. Biochem.* **63**, 777–822
13. Gregory, D.S., Martin, A.C.R. Cheetham, J.C., and Rees, A.R. (1993) The prediction and characterization of metal binding sites in proteins. *Protein Eng.* **6**, 29–35
14. Chakrabarti, P. (1990) Systematics in the interaction of metal ions with the main-chain carbonyl group in protein structures. *Biochemistry* **29**, 651–658
15. Lippard, S.J. and Berg, J.M (1994) *Principles of Bioinorganic Chemistry*, pp. 21–41, University Science Books, California

16. Nobeli, I., Laskowski, R.A., Valdar, W.S.J., and Thornton, J.M. (2001) On the molecular discrimination between adenine and guanine by proteins. *Nucleic Acids Res.* **29**, 4294–4309
17. Hayashi, H. (1995) Pyridoxal enzymes: mechanistic diversity and uniformity. *J. Biochem.* **118**, 463–473
18. Helenius, A. and Aebi, M. (2001) Intracellular Functions of N-Linked Glycans. *Science* **291**, 2364–2369