Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Integrated database of information from structural genomics experiments

Yukuhiko Asada,[a] Michihiro Sugahara,[a] Hisashi Mizutani,[a] Hisashi Naitow,[a] Tomoyuki Tanaka,[a] Yoshinori Matsuura,[a] Yoshihiro Agari,[b] Akio Ebihara,[b] Akeo Shinkai,[b] Seiki Kuramitsu,[b] Shigeyuki Yokoyama,[c] Eri Kaminuma,[d] Norio Kobayashi,[d] Koro Nishikata,[d] Sayoko Shimoyama,[d] Tetsuro Toyoda,[d] Tetsuya Ishikawa[a] and Naoki Kunishima[a]*

[a]Protein Crystallography Research Group, RIKEN SPring-8 Center, Harima Institute, 1-1-1 Kouto, Sayo-cho, Sayo-gun, Hyogo 679-5148, Japan, [b]SR System Biology Research Group, RIKEN SPring-8 Center, Harima Institute, 1-1-1 Kouto, Sayo-cho, Sayo-gun, Hyogo 679-5148, Japan, [c]Systems and Structural Biology Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan, and [d]Bioinformatics and Systems Engineering Division, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Correspondence e-mail: kunisima@spring8.or.jp

Received 29 November 2012
Accepted 17 January 2013

Information from structural genomics experiments at the RIKEN SPring-8 Center, Japan has been compiled and published as an integrated database. The contents of the database are (i) experimental data from nine species of bacteria that cover a large variety of protein molecules in terms of both evolution and properties (http://database.riken.jp/db/bacpedia), (ii) experimental data from mutant proteins that were designed systematically to study the influence of mutations on the diffraction quality of protein crystals (http://database.riken.jp/db/bacpedia) and (iii) experimental data from heavy-atom-labelled proteins from the heavy-atom database HATODAS (http://database.riken.jp/db/hatodas). The database integration adopts the semantic web, which is suitable for data reuse and automatic processing, thereby allowing batch downloads of full data and data reconstruction to produce new databases. In addition, to enhance the use of data (i) and (ii) by general researchers in biosciences, a comprehensible user interface, Bacpedia (http://bacpedia.harima.riken.jp), has been developed.

1. Introduction

Structural genomics is a 'big science' in the post-genomic era with the aim of establishing the structural basis of biology through an encompassing determination of protein structures (Burley, 2000). Since the launch of the International Structural Genomics Organization (ISGO; http://www.isgo.org) in 2001, a number of projects have contributed to a remarkable increase in the number of deposited macromolecular structures in the Protein Data Bank (PDB; http://www.rcsb.org/pdb). As of 2012, 27 structural genomics centres in eight countries are listed on the ISGO homepage. Among these, the three pioneers of structural genomics initiatives are the National Project on Protein Structural and Functional Analyses (NPPSFA, also known as the 'Protein 3000 Project') in Japan (http://mdbpr4.genes.nig.ac.jp/p3k/index.html.en; Yokoyama *et al.*, 2000), the Protein Structure Initiative (PSI) in the USA (http://www.nigms.nih.gov/Research/FeaturedPrograms/PSI; Terwilliger, 2000) and Structural Proteomics in Europe (SPINE; Heinemann, 2000; Stuart *et al.*, 2006). As of November 2012, 21% of the structures deposited in the PDB, excluding those with 100% identity in amino-acid sequence, are from structural genomics projects, indicating their substantial contribution to structural biology. It has been reported that proteins that share over 30% sequence identity have the same main-chain folding (Kryshtafovych *et al.*, 2009). Therefore, as a result of efforts in structural genomics, about three-quarters of current protein sequences contain a domain recognized by a known sequence profile, indicating a high success rate in prediction of protein folding from sequence information (Levitt, 2009).

At the RIKEN SPring-8 Center, Japan, large-scale crystallographic analyses of bacterial proteins have been performed as a contribution to the Protein 3000 Project using SPring-8 X-rays in combination with

a high-throughput platform for structural studies (Sugahara *et al.*, 2008; Iino *et al.*, 2008). Generally, a structural genomics project produces a huge amount of experimental data, including data from gene manipulation, protein production, crystallization and crystallographic analysis. Once these data have been properly compiled as an integrated database, then it will serve as a useful resource in both prospective and retrospective senses, for instance in the development of new methodologies and in the construction of infrastructures for structural studies (Gerstein, 2000). Since structural genomics projects consume huge budgets from taxpayer's money, the dissemination of the experimental data that are produced is strictly required by society to justify the cost. For a database from structural genomics, wide accessibility to unpublished data is also demanded so as to utilize them in methodological developments (Hol, 2000). In the USA, the publication of experimental data from structural genomics has been examined at several of the PSI centres. For instance, the Joint Center for Structural Genomics (JCSG; http://www.jcsg.org/prod/newscripts/home.html) publicly opened a traceable experimental history of targets that have been deposited in the PDB with a download service of raw diffraction images. However, preliminary data from targets without PDB depositions are not available and users have to log in to the JCSG server to download the diffraction images, indicating limited accessibility to general researchers. To benefit researchers in a broad range of life-science fields, here we have compiled and published the experimental data from the Protein 3000 Project (April 2002–March 2007) as a contribution to the Integrated Database Project (October 2007–March 2011) funded by the Japanese government. To our knowledge, this is the first fully accessible publication of experimental data, including preliminary results and raw diffraction images, from a structural genomics project.

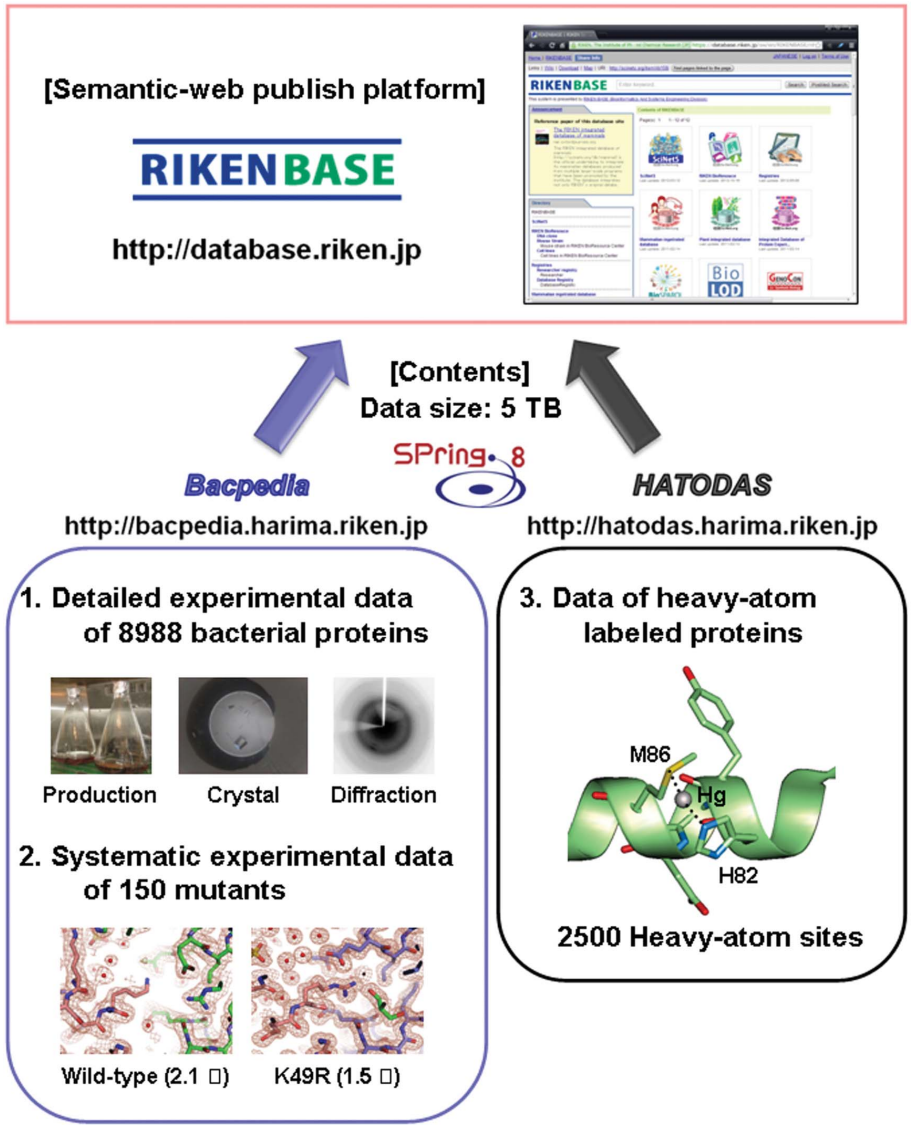

Figure 1
Publication of three databases from structural genomics experiments. The direct URL of the integrated database of protein experiments and structures on RIKENBASE is https://database.riken.jp/sw/en/Integrated_Database_of_Protein_Experiments_and_Structures/ria266i/.

2. Overall description

2.1. Original database

RIKEN has two major facilities for the determination of protein structures: the SPring-8 Center of Harima Institute for crystal structures and the NMR Center of Yokohama Institute for solution structures. These two facilities contributed equally to the Protein 3000 Project of Japan (http://mdbpr4.genes.nig.ac.jp/p3k/index.html.en). The Bioinformatics and Systems Engineering Division of RIKEN has a data platform called the RIKENBASE (http://database.riken.jp) for semantic web-based integration/publication of RIKEN databases (Masuya *et al.*, 2011). Using the RIKENBASE semantic web publication platform, we compiled, integrated and publicly opened the experimental information that had been produced in the Protein 3000 Project of protein crystallography at the RIKEN SPring-8 Center (Fig. 1). Since we had various formats of data at different laboratories in the SPring-8 campus, we first compiled the source data into a local database referred to as 'Bacpedia' (Supplementary Fig. S1[1]). Bacpedia is a Linux-based relational database that uses the SQL Server as database-administration software. The data unit of Bacpedia is constructed by IP-SAN (iSCSI) with 30 TB of disk space. It has two XML-based web user interfaces for search and editing. In the Integrated Database Project we also published the experimental data from the heavy-atom database HATODAS, which had been developed in the Protein 3000 Project (Sugahara *et al.*, 2005, 2009). We opened the integrated contents of the existing local database HATODAS so as to allow users outside RIKEN to download data for

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: MH5084). Services for accessing this material are described at the back of the journal.

reuse. The local database including Bacpedia and HATODAS before data integration is designated the 'original DB' (Fig. 2).

2.2. Complementary publication style

Experimental data compiled at the RIKEN SPring-8 Center are transferred and integrated onto the RIKENBASE platform and published from the same platform as a contribution to the Integrated Database Project (Fig. 1). Data integration onto the RIKENBASE platform denotes the reconstruction of an original DB using the RDF format (Resource Description Framework; http://www.w3.org/RDF), which is a standard format in semantic web technology (Berners-Lee *et al.*, 2001). This operation enables full data download in various formats including XML and TSV, and makes it possible for a research institute to establish a comprehensive data administration, for instance (Fig. 3*b*). Furthermore, integrated data retrieval using the *GRASE* (*General and Rapid Association Study Engine*; Kobayashi & Toyoda, 2008) search engine implemented at RIKEN leads to unexpected discoveries in life sciences (Fig. 3*d*). The semantic web technology of RIKENBASE has an advantage in data reuse and automation, thereby suiting researchers in bioinformatics. On the other hand, the original DBs Bacpedia and HATODAS implement user interfaces oriented to structural biologists and have an advantage in pinpoint searches for specific work (Fig. 2). According to user requests, we are taking a complementary publication style in which both the RIKENBASE and the original DBs cooperate to cover a wider range of researchers.

Full data downloading is available from the RIKENBASE, except for the diffraction-image data (Fig. 3*b*). Because of the large size of the files, the diffraction-image data are downloaded in fractional amounts from RIKENBASE (Fig. 3*c*). In the original Bacpedia, users can download diffraction data frame by frame using any downloader

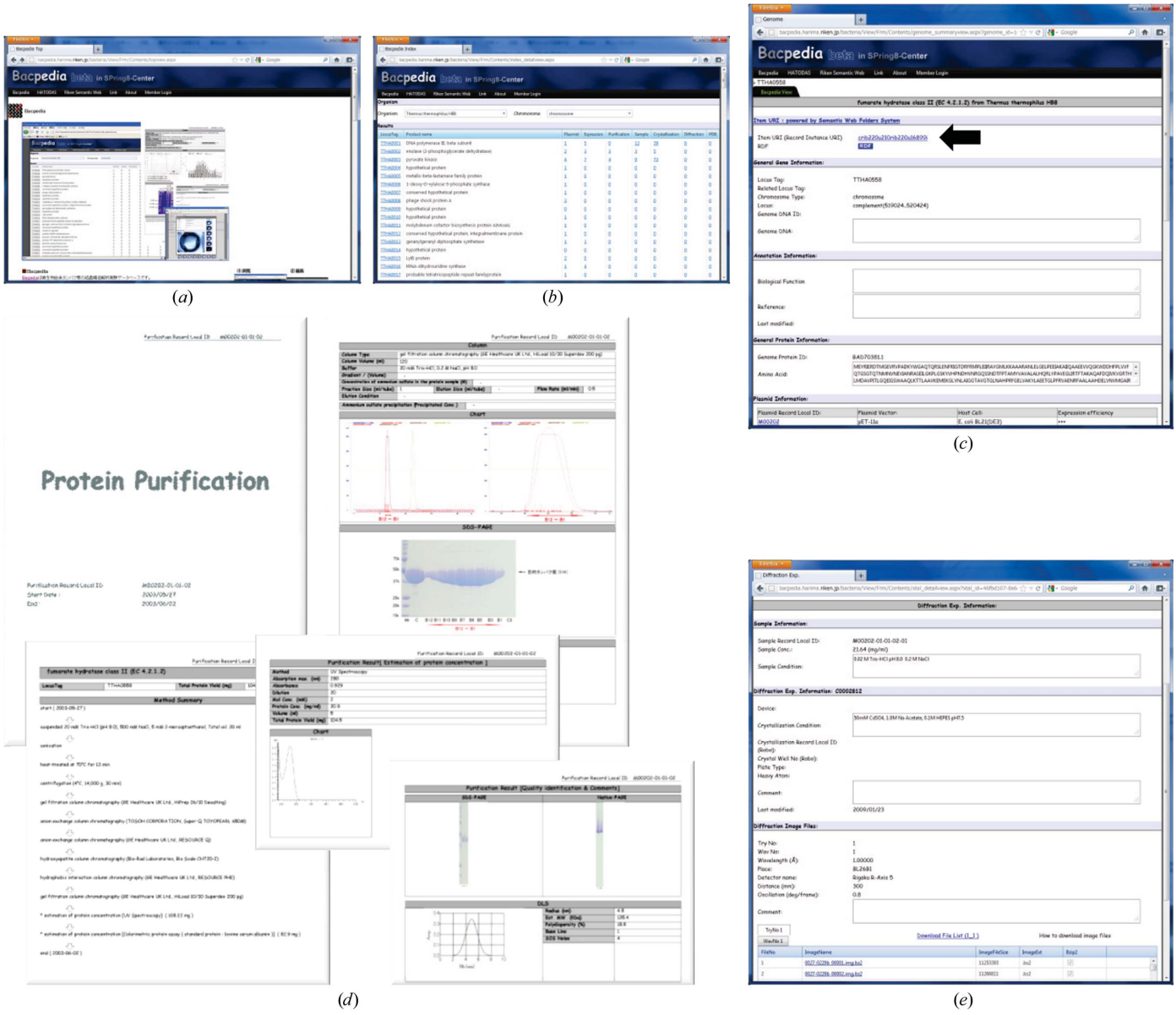

Figure 2
The original database Bacpedia (http://bacpedia.harima.riken.jp). (*a*) Home page. (*b*) Index view by protein ID. (*c*) Instance page, with a link to RIKENBASE indicated by an arrow. (*d*) Protein-purification report ready for printing. (*e*) Diffraction image download page.

Table 1
Contents of Bacpedia.

	Gene	Non-null
Mesophilic eubacterium		
Escherichia coli K-12	4341	3270 (75%)
Moderately thermophilic eubacterium		
Geobacillus kaustophilus HTA426	3540	335 (9%)
Extremely thermophilic eubacterium		
Thermus thermophilus HB8	2253	2062 (92%)
Hyperthermophilic eubacteria		
Aquifex aeolicus VF5	1551	374 (24%)
Thermotoga maritima MSB8	1846	275 (15%)
Hyperthermophilic archaea		
Pyrococcus horikoshii OT3	2065	1853 (90%)
Aeropyrum pernix K1	1700	177 (10%)
Methanocaldococcus jannaschii DSM2661	1770	226 (13%)
Sulfolobus tokodaii strain 7	2826	416 (15%)
Total	21892	8988 (41%)

specifying the URLs of images (Fig. 2*e*). As a condition of using the data, the CCO Creative Commons license (http://creativecommons.org) is employed, indicating that researchers can utilize data freely without any restriction like a public domain. Furthermore, login is no longer required to download all data, including diffraction images. Therefore, this database system should be useful to a wide range of researchers in life sciences.

2.3. Data update

The integrated database on RIKENBASE can be updated using the LinkData system (http://linkdata.org). For example, the N-terminal sequence data for 12 proteins from *Thermus thermophilus* HB8 were recently added to Bacpedia on RIKENBASE in July 2012 (http://linkdata.org/work/rdf1s139i). These data sets were based on a proteome analysis of *T. thermophilus* body extract by tandem mass spectrometry and will be helpful for structural studies. We are now ready to update this type of data in response to requests from users.

3. Contents of database

3.1. Proteins from bacteria (http://database.riken.jp/db/bacpedia)

The Protein Crystallography Research Group and the SR System Biology Research Group at the RIKEN SPring-8 Center contributed to the Protein 3000 Project through the X-ray crystallography of proteins mainly from microorganisms using the intense synchrotron radiation at SPring-8. In particular, for the hyperthermophilic eubacterium *Thermus thermophilus* HB8 structure determination has been completed for over 20% of a total of 2200 genes to date, making this bacterium one of the most analyzed organisms in terms of the three-dimensional structure of proteins.

In order to obtain the crystal structure of a protein, one has to follow several steps: (i) large-scale expression of the gene of interest, (ii) purification of the expressed protein, (iii) crystallization of the purified protein, (iv) an X-ray diffraction experiment for the protein crystals and (v) structure determination based on the diffraction data. A huge amount of information is produced at each stage of protein crystallography. Although these data had been collected vigorously in a few independent laboratories of RIKEN, their integrated publication had been hampered by differences in data formats. Therefore, to enhance the value of the data in terms of the reuse of information, the Protein Crystallography Research Group performed a data cleansing in which the original data in the RIKEN SPring-8 Center were edited and checked for data consistency and consolidated into the original database Bacpedia, in collaboration with the SR System Biology Research Group (Supplementary Fig. S1).

The experimental data from the X-ray crystallography of bacterial proteins would be useful for the structural analysis of homologous

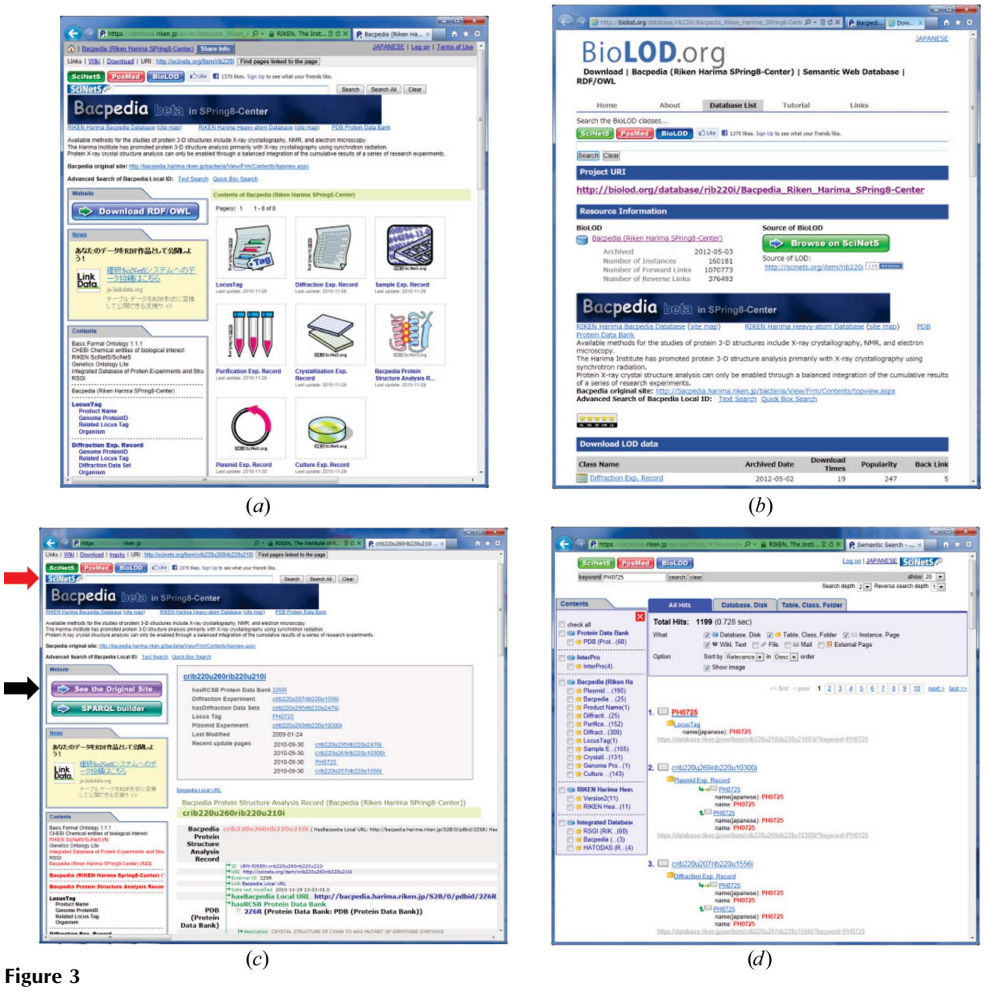

Figure 3
The integrated database Bacpedia on RIKENBASE (http://database.riken.jp/db/bacpedia). Similar interfaces are also available for HATODAS. (*a*) Home page. (*b*) Full data download page. (*c*) Semantic web instance page with a search tool and with a link to the original database, indicated by red and black arrows, respectively. (*d*) Integrated retrieval powered by the search engine *GRASE*.

proteins, for instance. In Bacpedia, we published all of the available experimental data, including those for preliminary/partial targets which had not reached structure determination. Thus, 8988 of a total of 21 892 entry genes (41%) have non-null experimental data, indicating high completeness of this database (Table 1). In both the original and the integrated Bacpedia, each instance page is linked together to enhance their complementary use (Figs. 2*c* and 3*c*). For bioinformatics researchers oriented towards data-mining studies, full data downloading from Bacpedia on RIKENBASE may be useful to process data in an automated way (Fig. 3*b*). On the other hand, in the case where a structural biologist is searching for information related to his/her target protein, the original Bacpedia (http://bacpedia.harima.riken.jp) may work better. A comprehensible protein-purification report for each protein is available in a format ready for printing (Fig. 2*d*). Furthermore, in the case of proteins from *T. thermophilus* HB8, a gene of interest is available from the RIKEN BioResource Center (http://www.brc.riken.jp) to produce an experimental strategy for the target protein through comparison with its homologue.

3.2. Mutant proteins (a part of http://database.riken.jp/db/bacpedia)

In order to study the stability, crystallization and heavy-atom derivatization of proteins, the Protein Crystallography Research Group of RIKEN performed extensive mutation experiments including crystal structure determination. The model proteins used were phosphoglycerate mutase from *T. thermophilus* HB8 (TTHB049) and diphthine synthase from the hyperthermophilic archaeon *Pyrococcus horikoshii* OT3 (PH0725).

TTHB049 is a small monomeric protein of 177 amino-acid residues which yields the steady production of a large amount of samples for crystallographic experiments. Another advantage of using this protein is its excellent applicability to thermostabilization studies of proteins by introducing mutations, as the denaturation temperature (T_m) of this protein is about 353 K (data not shown), which is not very high for *T. thermophilus*. The other model, PH0725, is a medium-sized dimeric protein with 530 amino-acid residues which also yields a steady production of samples for crystallographic study. The wild-type PH0725 protein provides crystals of medium diffraction quality at 2.1 Å resolution, with room for improvement (Kishishita *et al.*, 2008). Furthermore, most PH0725 mutants produce crystals that are isomorphous to the wild-type crystals. Thus, this protein is an excellent model to study the relationship between mutation and crystal quality. From a series of mutation experiments on PH0725, we showed that the diffraction quality of proteins can be improved by introducing mutations in the crystal-packing regions (Mizutani *et al.*, 2008).

In this work, we published experimental data from crystallography of 150 mutants of these two model proteins which would be useful for protein design based on data mining, for instance (Fig. 1). Researchers can download the full data from Bacpedia on RIKENBASE (Fig. 3*b*). Extraction of the mutant data can be performed by searching on the protein ID (for example, PH0725). These experiments on mutant proteins were performed using a standardized protocol, thereby providing more homogeneous and detailed data when compared with other proteins in Bacpedia.

3.3. Heavy-atom data (http://database.riken.jp/db/hatodas)

The heavy-atom database HATODAS was originally developed as software/a database in protein crystallography to support the heavy-atom-derivatization process of a target protein (http://hatodas.harima.riken.jp; Sugahara *et al.*, 2005). HATODAS is a database of known heavy-atom-derivatized proteins that suggests potential heavy-atom reagents for derivatization experiments based on the amino-acid sequence of the target protein and its crystallization conditions. The latest version (v.2) of HATODAS expanded its entries by about fourfold and added several new functions including potentiality scoring to prioritize the heavy-atom reagents suggested for experiments (Sugahara *et al.*, 2009).

In this work, we made available publicly the experimental data for 2500 heavy-atom-labelled proteins from the original HATODAS as part of the integrated database on RIKENBASE to allow full data download and data reuse, which would facilitate the development of new technologies in protein engineering, for instance (Fig. 1). Researchers can download the full data from HATODAS on RIKENBASE for data mining in the same way as Bacpedia (Fig. 3*b*).

4. Discussion

4.1. Status of utilization

The integrated Bacpedia/HATODAS on RIKENBASE and the original Bacpedia were published on 23 July 2009 and 6 July 2010, respectively. In this work, we published the data from structural genomics experiments of bacterial proteins consisting of protein production (11 800 plasmid constructions, 7700 expressions and 3600 purifications), crystallization (12 500 000 observation images) and diffraction experiments (700 data sets), and the experimental data for 2500 heavy-atom labelled proteins. Based on a statistical analysis as of November 2012, the original Bacpedia has mainly been accessed by users inside Japan, with about 9100 page views since its publication. Therefore, we believe that international utilization of this database will be enhanced in the future. In contrast, the original HATODAS has been accessed from many countries, with about 35 800 page views since the publication of the integrated version. In addition to the 31% of accesses from Japan, it is accessed from the USA (18%), Germany (9%), China (8%), the UK (6%) and so on, indicating worldwide recognition.

4.2. Future applications

The expected availability of the three published databases is described below.

In the database of bacterial proteins, detailed experimental data in protein crystallography are systematically collected with excellent accessibility. Because it includes preliminary/partial data, users can utilize all the information from nine species of bacteria covering large variety of protein molecules in terms of both evolution and properties (Table 1). Many other applications apart from use as reference information for structural studies are conceivable. For instance, on the basis of results from data mining, researchers can develop effective software for structure determination of proteins (Hol, 2000). Furthermore, it can be used as a representative benchmark set for the evaluation of various data in biosciences, making a basic contribution to broader society.

A concrete example of an application of the database of mutant proteins is the development of high-precision homology modelling. Using the current technology of homology modelling, which predicts a model for the target protein from the crystal structure of a homologue, it is generally difficult to utilize the low-precision homology model in pharmaceutical development (Novotný *et al.*, 1984). In our database, structures of systematic mutant proteins in the same crystal form are given, indicating the possibility of improving the algorithm of homology modelling on the basis of structures. Development of

this new technology would contribute to the structural prediction of target proteins from homologous structures.

Using the database of heavy-atom-labelled proteins, developments in protein engineering are expected. To date, we have found many motif sequences that bind heavy atoms specifically (Sugahara *et al.*, 2005). Thus, it is possible to label a target protein with heavy-atom reagents by introducing a heavy-atom-binding motif using a protein-engineering technique, which can be used in various applications.

5. Conclusion

The semantic web is a key technology in the integrated database. Since it allows data reuse, which can lead to unexpected discoveries, various applications including pharmaceutical development are beginning (Cannata *et al.*, 2008; Splendiani, 2008; Frey, 2009; Wild *et al.*, 2012). This work is regarded as a model case of database integration that will provide the basis of heuristic bioinformatics by creating an international network with other databases worldwide in the future.

We thank the RIKEN staff in the Bioinformatics and Systems Engineering Division and the RIKEN SPring-8 Center for their help in database construction and thank Dr T. Kumarevel of the RIKEN SPring-8 Center for proofreading the manuscript. This work was supported by the Integrated Database Project funded by the Ministry of Education, Culture, Sports, Science and and Technology (MEXT) of Japan. The original databases Bacpedia and HATODAS are partially supported by the Platform for Drug Discovery, Informatics, and Structural Life Science from the MEXT, Japan. The RIKEN-BASE is partially supported by the Database Integration Program of the National Bio Science Center of the Japan Science and Technology Agency.

References

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). *Sci. Am.* **284**, 34–43.
Burley, S. K. (2000). *Nature Struct. Biol.* **7**, Suppl., 932–934.
Cannata, N., Schröder, M., Marangoni, R. & Romano, P. (2008). *BMC Bioinformatics*, **9**, Suppl. 4, S1.
Frey, J. G. (2009). *Drug Discov. Today*, **14**, 552–561.
Gerstein, M. (2000). *Nature Struct. Biol.* **7**, Suppl., 960–963.
Heinemann, U. (2000). *Nature Struct. Biol.* **7**, Suppl., 940–942.
Hol, W. G. (2000). *Nature Struct. Biol.* **7**, Suppl., 964–966.
Iino, H., Naitow, H., Nakamura, Y., Nakagawa, N., Agari, Y., Kanagawa, M., Ebihara, A., Shinkai, A., Sugahara, M., Miyano, M., Kamiya, N., Yokoyama, S., Hirotsu, K. & Kuramitsu, S. (2008). *Acta Cryst.* F**64**, 487–491.
Kishishita, S., Shimizu, K., Murayama, K., Terada, T., Shirouzu, M., Yokoyama, S. & Kunishima, N. (2008). *Acta Cryst.* D**64**, 397–406.
Kobayashi, N. & Toyoda, T. (2008). *Bioinformatics*, **24**, 1002–1010.
Kryshtafovych, A., Fidelis, K. & Moult, J. (2009). *Proteins*, **77**, 217–228.
Levitt, M. (2009). *Proc. Natl Acad. Sci. USA*, **106**, 11079–11084.
Masuya, H. *et al.* (2011). *Nucleic Acids Res.* **39**, D861–D870.
Mizutani, H., Saraboji, K., Malathy Sony, S. M., Ponnuswamy, M. N., Kumarevel, T., Krishna Swamy, B. S., Simanshu, D. K., Murthy, M. R. N. & Kunishima, N. (2008). *Acta Cryst.* D**64**, 1020–1033.
Novotný, J., Bruccoleri, R. & Karplus, M. (1984). *J. Mol. Biol.* **177**, 787–818.
Splendiani, A. (2008). *BMC Bioinformatics*, **9**, Suppl. 4, S6.
Stuart, D. I., Jones, E. Y., Wilson, K. S. & Daenke, S. (2006). *Acta Cryst.* D**62**, doi:10.1107/S0907444906024759.
Sugahara, M., Asada, Y., Ayama, H., Ukawa, H., Taka, H. & Kunishima, N. (2005). *Acta Cryst.* D**61**, 1302–1305.
Sugahara, M., Asada, Y., Shimada, H., Taka, H. & Kunishima, N. (2009). *J. Appl. Cryst.* **42**, 540–544.
Sugahara, M. *et al.* (2008). *J. Struct. Funct. Genomics*, **9**, 21–28.
Terwilliger, T. C. (2000). *Nature Struct. Biol.* **7**, Suppl., 935–939.
Wild, D. J., Ding, Y., Sheth, A. P., Harland, L., Gifford, E. M. & Lajiness, M. S. (2012). *Drug Discov. Today*, **17**, 469–474.
Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Terada, T., Ito, Y., Matsuo, Y., Kuroda, Y., Nishimura, Y., Kyogoku, Y., Miki, K., Masui, R. & Kuramitsu, S. (2000). *Nature Struct. Biol.* **7**, Suppl., 943–945.