Published online 14 November 2008

Nucleic Acids Research, 2009, Vol. 37, Database issue **D365–D368**
doi:10.1093/nar/gkn790

The protein structure initiative structural genomics knowledgebase

Helen M. Berman[1,*], John D. Westbrook[1], Margaret J. Gabanyi[1], Wendy Tao[1], Raship Shah[1], Andrei Kouranov[1], Torsten Schwede[2], Konstantin Arnold[2], Florian Kiefer[2], Lorenza Bordoli[2], Jürgen Kopp[2,3], Michael Podvinec[2], Paul D. Adams[4], Lester G. Carter[4], Wladek Minor[5], Rajesh Nair[6] and Joshua La Baer[7]

[1]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, [2]Swiss Institute of Bioinfomatics & Biozentrum, University of Basel, CH-4056 Basel, Switzerland, [3]Biochemie-Zentrum, Heidelberg University, D-69120 Heidelberg, Germany, [4]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [5]Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908-0736, [6]Department of Biochemistry & Molecular Biophysics, Columbia University, New York, NY 10027 and [7]Harvard Institute of Proteomics & Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

Received September 16, 2008; Accepted October 8, 2008

ABSTRACT

The Protein Structure Initiative Structural Genomics Knowledgebase (PSI SGKB, http://kb.psi-structuralgenomics.org) has been created to turn the products of the PSI structural genomics effort into knowledge that can be used by the biological research community to understand living systems and disease. This resource provides central access to structures in the Protein Data Bank (PDB), along with functional annotations, associated homology models, worldwide protein target tracking information, available protocols and the potential to obtain DNA materials for many of the targets. It also offers the ability to search all of the structural and methodological publications and the innovative technologies that were catalyzed by the PSI's high-throughput research efforts. In collaboration with the Nature Publishing Group, the PSI SGKB provides a research library, editorials about new research advances, news and an events calendar to present a broader view of structural biology and structural genomics. By making these resources freely available, the PSI SGKB serves as a bridge to connect the structural biology and the greater biomedical communities.

INTRODUCTION

The goal of the worldwide structural genomics initiative is to determine the 3D structures of proteins on a genomic scale. Since 2001, these efforts have resulted in more than 6772 structure depositions to the PDB (1), 3251 of which are from the National Institutes of Health-sponsored Protein Structure Initiative Centers. In order to determine these structures in a high-throughput manner, the PSI Centers have developed advanced technologies to facilitate these processes, including protein production, crystallization, structure determination, refinement and analysis.

The Protein Structure Initiative Structural Genomics Knowledgebase (PSI SGKB) (http://kb.psi-structuralgenomics.org) was launched in early 2008 with the goal of making the results of the PSI initiative widely available to the broad community of biologists (2). The PSI SGKB provides access to the annotated protein structures, the models that can be leveraged from them, associated functional predictions, experimental protocols and tracking information. The new resource integrates this structural information with relevant scientific information from external data resources in a coherent and contextual format. The PSI SGKB also provides the descriptions of a vast array of technologies, protein production protocols and software applications. By making all of these products accessible to the greater community, it is anticipated that the PSI SGKB will become an enabling resource for biologists, biochemists, functional genomicists, pharmacologists, educators and physicians.

ARCHITECTURE

The PSI SGKB is a portal that provides integrated access to the PSI Centers, external biological databases, the PDB

*To whom correspondence should be addressed. Tel: +1-732-445-0103; Fax: +1-732-445-4320; Email: berman@rcsb.rutgers.edu and a set of widely distributed resources, some of which are portals themselves. These resources include:

Experimental data tracking: TargetDB (http://targetdb.rcsb.org) (3) and PepcDB (http://pepcdb.rcsb.org) (4) were established to track the progress of targets that are being worked on by the worldwide structural genomics centers. TargetDB gives the status of each target and PepcDB provides information about the protocols used for protein production and the reasons for stopping work on any one target. Data are regularly collected, tracked and made available via the TargetDB and PepcDB web sites. These resources have query and report functionality. PepcDB now provides access to the general protocols for the most successful experimental trial as well as any specific details associated with the trial. Cross-references are provided to the Research Collaboratory for Structural Biology (RCSB) PDB (5), Pfam (6), Superfamily (7), TIGR Families (8), ProDom (9), iProClass (10) and Prosite (11). Sequence family (e.g. BIG and MEGA assignments) and target classification details (e.g. biomedical, community-nominated) are now collected from PSI centers. As PepcDB grows, this resource will become an indispensable and truly unique resource for biologists who are expressing and purifying proteins for their own experiments.

Materials repository: The PSI Materials Repository (PSI MR) (http://www.hip.harvard.edu/PSIMR/index.htm) has been established at Harvard University. A mechanism for storing and distributing clones is in place. Material transfer agreements have been established between the PSI centers and the PSI MR. The specifications for the data to accompany physical samples have been set. These data include the information required to ensure the interoperability of this repository with TargetDB and PepcDB. The information collected for each PSI clone will also be stored in a searchable database. Using this database, researchers will be able to select and order clones online for a minimal fee that covers processing, handling and shipping.

Homology modeling: For every structure determined by the PSI Centers, hundreds of models could be made using a variety of established methods. This has been done by all of the PSI centers. At a workshop held in 2005 at Rutgers University, it was proposed that a portal for models should be launched (12). This would allow access to a variety of models predicted by different methods for any target protein. The Protein Modeling Portal [PMP; http://www.proteinmodelportal.org; (13)] has been established at the Swiss Institute of Bioinformatics (Biozentrum University of Basel) and is headed by Torsten Schwede. The PMP currently provides access to several million prebuilt models from the four PSI centers and publicly available model databases [i.e. ModBase (14) and SWISS-MODEL Repository (15)]. The PMP can be accessed from the PSI SGKB through web service queries, or by directly searching the portal for models of specific or similar protein sequences or models built on specific template structures. The PSI SGKB links to the PMP pages which display information on individual models (or sets of models), as well as functional annotation of the target protein sequence.

Annotation: Many different annotations are possible for every target, including structure determination and validation details; sequence information, including possible family and domain assignments; structure information, including surface characteristics; cavities; potential and actual active sites; fold classifications; protein–protein and protein–ligand interactions; and structure–function relationships. The PSI centers have created services that provide access to many of these annotations. The PSI SGKB links to the PSI interactive services, summaries and galleries of annotation information, with several of these resources integrated directly into PSI SGKB search reports. In addition, the search reports link to approximately 50 additional annotation resources of sequence, structure and function such as UniProt (16), National Center for Biotechnology Information (NCBI) (17,18), Class, Architecture, Topology, and Homologous Superfamily (CATH) (19), Structural Classification of Proteins (SCOP) (20), and Gene Ontology (GO) (21). A complete list of annotation sources is maintained on an Annotation Resources page (http://kb.psi-structuralgenomics.org/KB/annotation-resources.html). The 'Workshop on the Biological Annotation of Novel Proteins' (7 and 8 March 2008; http://annotation-workshop.rutgers.edu/) was convened to collect detailed recommendations and requirements for future annotation information to be incorporated into the PSI SGKB.

Technology development: Each PSI center has developed a variety of cutting edge technologies for all stages of the structure determination pipeline (22). These technologies have been critical to the success of the PSI program. Additionally, the descriptions of these technologies are an important resource for the broader biological community for use in other research. Established at Lawrence Berkeley Laboratory under the leadership of Paul Adams, the Technology Portal (https://isswprod.lbl.gov/PSIKBPortal/) currently provides access to summaries of key PSI technologies with links to the responsible PSI center and/or related publication. Information within the Technology Portal is accessible through keyword searches of the PSI SGKB.

Metrics: A quantitative assessment of the productivity of the PSI Centers is available via metrics, such as the numbers of distinct and novel structures. The metrics were articulated in a PSI Steering Subcommittee on Goals and Milestones report (http://targetdb.rcsb.org/Metrics/Milestones.html). Tabulations of key metrics from these recommendations are updated regularly at http://targetdb.rcsb.org/Metrics/SummaryTable.html and http://targetdb.rcsb.org/Metrics/MilestonesTables.html.

Publications: All articles published by PSI scientists are collected into a central Publications Resource (http://olenka.med.virginia.edu/psi) developed by Wladek Minor at the University of Virginia. Lists of citations are categorized as structural or methodological, and include the PubMed identifier and the number of times the article has been cited. With this final piece of information, the resource calculates and tracks the following (with current values at the time this manuscript was prepared): total number of articles published by all Centers (1036), number of articles cited more than five times (561), total

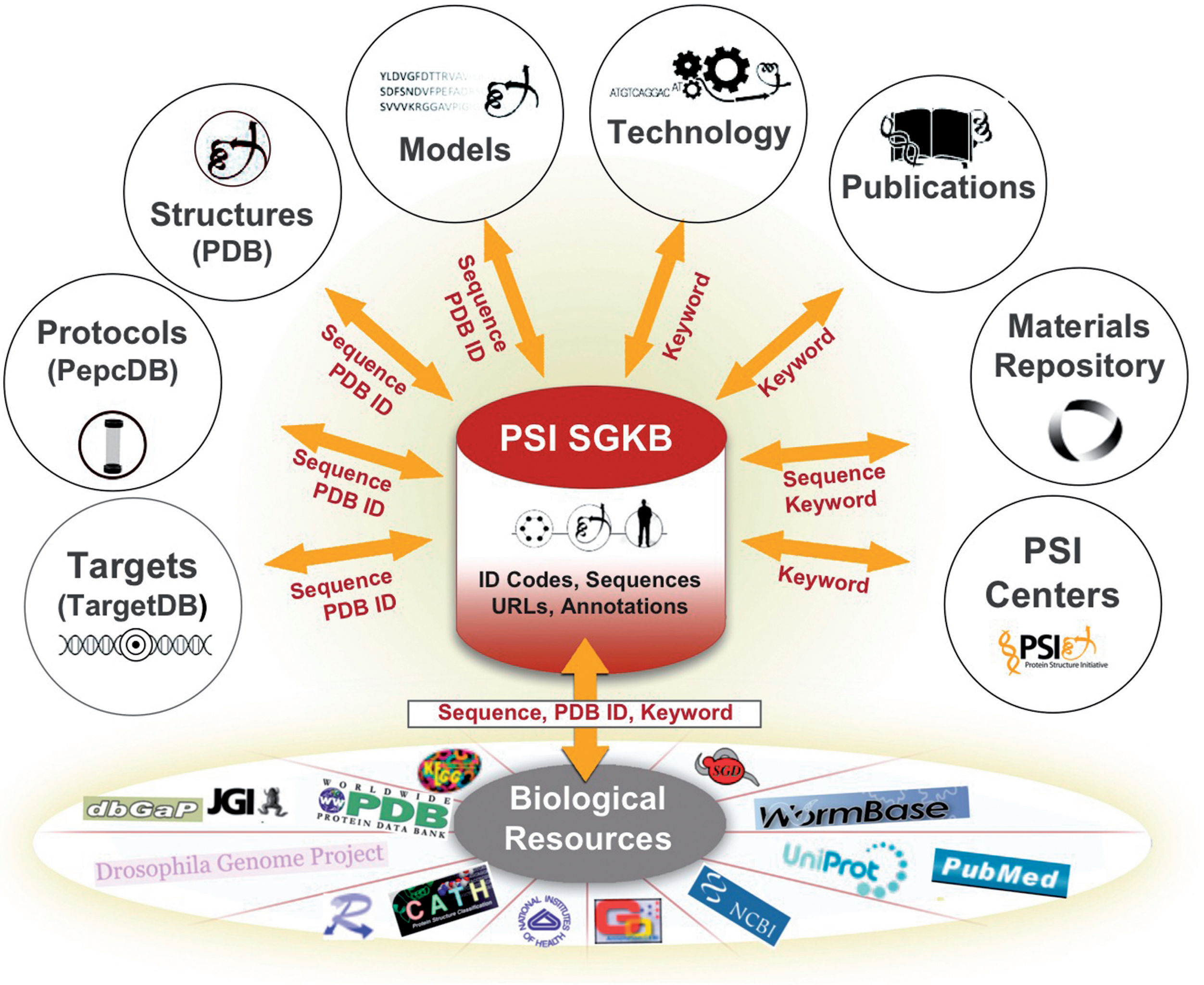

Figure 1. A functional view of the PSI SGKB resource. The PSI SGKB portal database is comprised of ID codes, sequences, external URLs and annotations. PDB ID, sequence, or keyword queries made to the PSI SGKB access other PSI data portals (TargetDB, PepcDB, PMP, Technology Portal, Publications Resource, MR, PSI Center websites), PDB and 50 other external biological resources (a selection is shown).

number of citations (16 077), average number of times a structural (10.7) or methodological (20.5) paper has been cited and total impact of the articles by journal impact rating (4985). These values are also available by individual PSI center. The resource additionally charts the number of publications by year and by impact factor. These statistics are essential to track the impact of the PSI efforts, since outreach and the dissemination of PSI-catalyzed research are central to the PSI mission.

SEARCH CAPABILITIES

Integrating information from each PSI center with external resources is a key focus of the PSI SGKB. This integration already exists in resources like the RCSB PDB, which contain annotations about folds from CATH (19) and SCOP (20), function via Gene Ontology (21) and sequence via UniProt (16). TargetDB and PepcDB are cross-linked to sequence family and domain databases. The PSI SGKB takes this integration a large step further. In the current PSI SGKB release, it is possible to perform a *single query* to extract information from the PDB, PepcDB, TargetDB, the PMP, the Technology Portal, the Publications Resource, the MR and other biological annotation resources (Figure 1). For example, a user can type in a sequence or a PDB ID code to extract reports containing structural information and annotations, protein production protocols, predicted models, available DNA clone materials and related target status, without having to go to each individual module. Keyword searches query technology information, publication information (including title and abstracts for structural genomics-related target and structure publications), the MR and all indexed pages from the PSI Centers.

THE NPG PSI SGKB GATEWAY

The scope and potential outreach of the PSI SGKB was expanded recently when it joined forces with the Nature Publishing Group to create a PSI SGKB Gateway. The Gateway enhances the PSI SGKB's powerful query feature with articles and resources that highlight research findings, new technologies, and general structural biology news. A *Research Library* catalogs articles relevant to structural biology and genomics. A description of a particularly interesting molecule is featured monthly. The *Functional Sleuth* section presents information about molecules of unknown function to challenge biologists to provide further insights about these molecules. News alerts and Really Simple Syndication (RSS) feeds will help alert the scientific community about the progress of structural genomics.

FUTURE DEVELOPMENTS

The core resources for the PSI SGKB are in place. New developments include the construction of a pipeline to collect and calculate an expanded set of annotations. A simplified matrix presentation will be provided for the new annotations and will highlight the structures requiring further functional characterization. Another planned capability will be the use of data mining to fully exploit these annotations.

FUNDING

National Institute of General Medical Sciences. Funding for open access charge: The PSI SGKB is supported by the National Institutes of Health (NIH) as a sub-grant under Prime Agreement Award Number: 3U54GM074958-04S1.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
2. Berman,H. (2008) Harnessing knowledge from structural genomics. *Structure*, **16**, 16–18.
3. Chen,L., Oughtred,R., Berman,H.M. and Westbrook,J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, **20**, 2860–2862.
4. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305.
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. Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.
7. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, D308–D313.
8. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
9. Corpet,F., Gouzy,J. and Kahn,D. (1998) The ProDom database of protein domain families. *Nucleic Acids Res.*, **26**, 323–326.
10. Wu,C.H., Xiao,C., Hou,Z., Huang,H. and Barker,W.C. (2001) iProClass: an integrated, comprehensive and annotated protein classification database. *Nucleic Acids Res.*, **29**, 52–54.
11. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
12. Berman,H.M., Burley,S.K., Chiu,W., Sali,A., Adzhubei,A., Bourne,P.E., Bryant,S.H., Dunbrack,R.L. Jr., Fidelis,K., Frank,J. *et al.* (2006) Outcome of a workshop on archiving structural models of biological macromolecules. *Structure*, **14**, 1211–1217.
13. Arnold,K., Kiefer,F., Kopp,J., Battey,J.N., Podvinec,M., Westbrook,J., Berman,H., Bordoli,L. and Schwede,T. (2008) The protein model portal. *J. Struct. Funct. Genomics* (in press).
14. Pieper,U., Eswar,N., Davis,F.P., Braberg,H., Madhusudhan,M.S., Rossi,A., Marti-Renom,M., Karchin,R., Webb,B.M., Eramian,D. *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.
15. Kopp,J. and Schwede,T. (2006) The SWISS-MODEL repository: new features and functionalities. *Nucleic Acids Res.*, **34**, D315–D318.
16. The UniProt Consortium. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
17. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
18. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **36**, D13–21.
19. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
20. Conte,L., Bart,A., Hubbard,T., Brenner,S., Murzin,A. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
21. The UniProt Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
22. Perkel,J. (2008) Structural—proteomics: the relentless pursuit of protein shape. *Science*, **321**, 707–710.