Spring 2008 • Number 37

RCSB PDB PROTEIN DATA BANK

Published quarterly by the Research Collaboratory for Structural Bioinformatics Protein Data Bank

# NEWSLETTER

Weekly RCSB PDB news is available online at www.pdb.org

## Contents



**SNAPSHOT: APRIL 1, 2008**

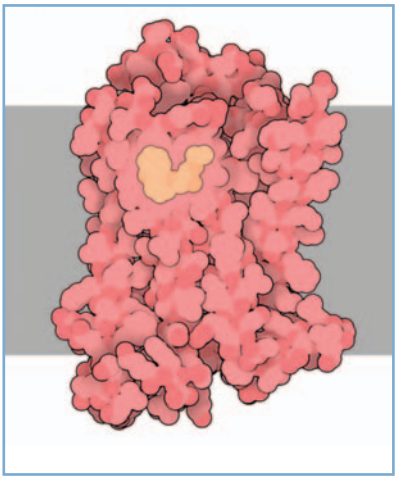
49760 released atomic coordinate entries

| MOLECULE TYPE | | EXPERIMENTAL TECHNIQUE | |
|---|---|---|---|
| 45906 | proteins, peptides, and viruses | 42342 | X-ray |
| 1839 | nucleic acids | 7150 | NMR |
| 1982 | protein/nucleic acid complexes | 170 | electron microscopy |
| 33 | other | 98 | other |
| | | 31499 | structure factor files |
| | | 3931 | NMR restraint files |

**Participating RCSB Members:**
**Rutgers • SDSC/SKAGGS/UCSD**
E-mail: info@rcsb.org
Web: www.pdb.org • FTP: ftp.wwpdb.org

The RCSB PDB is a member of the wwPDB (www.wwpdb.org)

## Message from the RCSB PDB



*April 2008 feature on adrenergic receptors marks the 100th Edition of* Molecule of the Month

April 1 marked the **100th edition** of the *Molecule of the Month*, a series produced by David S. Goodsell and featured on the RCSB PDB website.

Since January 2000, this series has explored the structure and function of proteins and nucleic acids found in the PDB archive such as transfer RNA, anthrax toxin, and multidrug resistance transporters. To commemorate this event, the RCSB PDB will be offering temporary tattoos of an adrenergic receptor at upcoming meetings. The feature is also available in a specially formatted PDF.

Written and illustrated by David S. Goodsell (The Scripps Research Institute), the *Molecule of the Month* provides an easy introduction to the RCSB PDB for teachers and students. It is used in many classrooms to introduce structures to students, and is an integral part of the protein modeling event at the Science Olympiad.

The text and images are related to the featured molecule; the RCSB PDB pages link to examples of the molecule. In response to requests, a view of the highlighted structure in Jmol is included in new features to provide an interactive view of the molecule.

New *Molecule of the Month* features are made available from the RCSB PDB home page with the first update of each month. Alphabetical and chronological listings of past issues are provided. wwPDB partner PDBj has recently started to translate the Molecule of the Month into Japanese.

Links to the series are also available from RCSB PDB's Structure Explorer pages. Selecting "Learn more: [M]" takes the reader to any *Molecule of the Month* feature related to that particular entry.

*David S. Goodsell*

To create the series, Goodsell combines his artistic talent with his scientific expertise in his visual representations of molecular biology. He creates his images so as to capture his excitement about science and communicate it to others.

"The combination of art and science gives me a way to access the wonder of nature. It makes me really look at results and think about them in a deeper way," Goodsell says. "The thing that drives me continually is the beauty of these objects that I'm working on and being amazed at how unusual they are."

# Data Deposition and Processing

## sf-convert: A Format Conversion Tool for Structure Factor Files

The command-line program, sf-convert, can easily translate data in various formats to the mmCIF format for use with ADIT validation and deposition software. sf-convert can also translate structure factors already released in the PDB from mmCIF to different formats.

This tool can input files from the following programs and formats: mmCIF, CIF, MTZ, CNS, Xplor, HKL2000, Scalepack, Dtrek, TNT, SHELX, SAINT, EPMR, XSCALE, XPREP, XTALVIEW, X-GEN, XENGEN, MULTAN, MAIN, and OTHER (an ASCII file with H, K, L, F, and SigmaF separated by a space).

sf-convert can then output the data formatted as mmCIF, MTZ, CNS, TNT, SHELX, EPMR, XTALVIEW, HKL2000, Dtrek, XSCALE, MULTAN, MAIN, or OTHER.

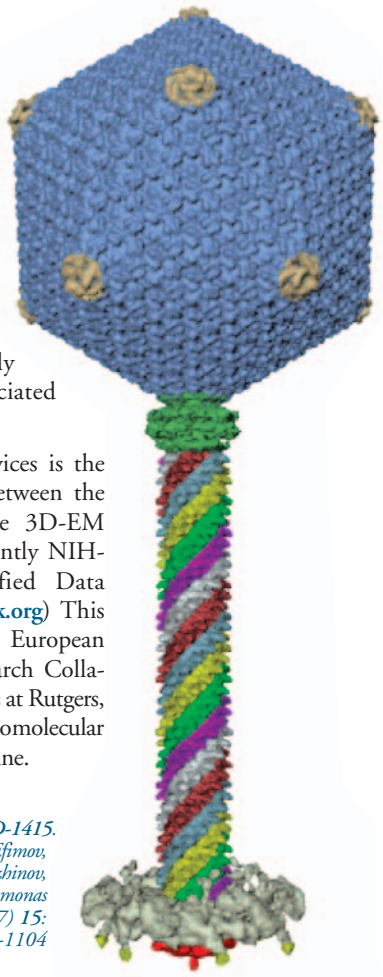
sf-convert is available for download from **sw-tools.pdb.org**.

## EmDep2: Deposit EM Maps at the MSD-EBI or RCSB PDB

Electron microscopy map data can now be deposited to the Electron Microscopy Data Bank (EMDB) using the improved web-based tool EmDep2. EmDep2 is available from the existing deposition site at the MSD-EBI in Europe and also from a new deposition site at the RCSB PDB in the USA.

The EMDB contains experimentally determined 3D maps and associated experimental data and files.

This improvement to EMDB services is the first product of a collaboration between the European Network of Excellence 3D-EM (**www.3dem-noe.org**) and the recently NIH-funded Partnership for a Unified Data Resource for CryoEM (**emdatabank.org**) This partnership is comprised of the European Bioinformatics Institute, the Research Collaboratory for Structural Bioinformatics at Rutgers, and the National Center for Macromolecular Imaging at Baylor College of Medicine.



*EMD-1415.*
*A. Fokine, A. Battisti, V. Bowman, A. Efimov, L. Kurochkina, P. Chipman, V. Mesyanzhinov, M. Rossmann Cryo-EM Study of the Pseudomonas Bacteriophage phiKZ* Structure *(2007) 15: 1099-1104*

EMDB: **www.ebi.ac.uk/msd-srv/docs/emdb**
EmDep2 (EBI): **www.ebi.ac.uk/msd-srv/emdep**
EmDep2 (RCSB PDB): **emdb.rutgers.edu/emdep**

## 2008 Deposition Statistics

In the first quarter of 2008, 1626 experimentally-determined structures were deposited to the PDB archive.

The entries were processed by wwPDB teams at the RCSB PDB, MSD-EBI, and PDBj. Of the structures deposited in the first quarter of 2008, 75% were deposited with a release status of "hold until publication"; 22.5% were released as soon as annotation of the entry was complete; and 2.5% were held until a particular date.

90% of these entries were determined by X-ray crystallographic methods; 9% were determined by NMR methods. 97% of these depositions were deposited with experimental data. As of February 1, 2008, the deposition of experimental data is required.

During the same period of time, 1915 structures were released into the archive.

## Data Processing Versioning Procedures

Data in the PDB archive currently follow either PDB File Format Version 3.0 or 3.1. This is indicated in REMARK 4 of the file.

Version 3.0 is the format used for files released as a result of the Remediation Project.

Since August 1, 2007, all files processed and released into the archive have followed Version 3.1. When modifications have been made to files released prior to that date, they have been then re-released in Version 3.1.

Version 3.1 differs from Version 3.0 in descriptions of the biological unit (REMARK 300/350), geometry (REMARK 500), atom/residues modeled as zero occupancy (REMARK 475/480), non-polymer residues with missing atoms (REMARK 610), and metal coordination (REMARK 620). Documentation describing the differences between these versions is available at **www.wwpdb.org/docs.html**.

Since the beginning of March 2008, the REVDAT record indicates when a Version 3.0 file is re-released as Version 3.1 with the name "VERSN."

For example, if the journal record has been updated in an entry that previously followed Version 3.0, the REVDAT would appear as:

```
REVDAT 1 04-MAR-08 1ABC 1 JRNL VERSN
REVDAT 1 13-FEB-07 1ABC 0
```

There is no change to how depositors submit their files. Any required changes in nomenclature can be made automatically by the wwPDB during the annotation process.

Documentation about file formats and the Remediation Project is available at **www.wwpdb.org**.

# Data Query, Reporting, and Access

## Website Statistics

Website access statistics for the first quarter of 2008 are given below.

| MONTH | UNIQUE VISITORS | NUMBER OF VISITS | BANDWIDTH |
|---|---|---|---|
| JAN 08 | 128781 | 319459 | 426.87 GB |
| FEB 08 | 139444 | 338946 | 567.18 GB |
| MAR 08 | 152264 | 361999 | 642.98 GB |

## PDB Statistics

Which journal has published the most structures? What types of structures have been solved by more than one experimental method? Answers to these questions can be found by exploring the various statistics about the data in the PDB archive available by clicking the **PDB Statistics** link at the top of every page on the RCSB PDB website.

Charts, graphs, and tables related to content distribution include:

- *Summary Table of Released Entries*: Current PDB holdings grouped by experimental method and molecule type
- *Status of Unreleased Entries*: A pie chart that illustrates the status of unreleased entries
- Interactive histograms showing the archive by function, resolution, space group, source organism, journal, molecular weight, and enzyme classification
- Histogram showing the number of structures solved by structural genomics structures
- Table of proteins solved by multiple experimental methods
- Current statistics on redundancy in the archive

The growth in the number of structures released in the PDB archive can be seen per year, by experimental method, and by molecule type. Other graphs show the growth of unique protein classifications as defined by SCOP (**scop.mrc-lmb.cam.ac.uk/scop/index.html**) and CATH (**cath-www.biochem.ucl.ac.uk**).

## Time-stamped Copies of PDB Archive Available via FTP

A time-stamped snapshot of the PDB archive (**ftp.wwpdb.org**) as of January 7, 2008 has been added to **ftp://snapshots.rcsb.org/**.

Snapshots of the PDB have been archived annually since 2004. It is hoped that these snapshots will provide readily identifiable data sets for research on the PDB archive.

The script at **ftp://snapshots.rcsb.org/rsyncSnapshots.sh** may be used to make a local copy of a snapshot or sections of the snapshot.

The directory 20080107 includes the 48,161 experimentally-determined coordinate files that were current as of January 7, 2008. Coordinate data are available in PDB, mmCIF, and XML formats. The date and time stamp of each file indicates the last time the file was modified.

# Outreach and Education

## RCSB PDB Celebrates Teaching, Learning, and More

*The RCSB PDB exhibits at meetings to get feedback from our users about the resource.*

Recent education and outreach activities have included:

• Annotators made models of virus structures with local middle school students as part of Princeton University's Science and Engineering Expo on March 19. The models included marshmallow and toothpick representations of the viral shell and paper models of the dengue fever virus.

• An exhibit booth was also held at the Teaching & Learning Celebration in New York, NY, March 7-8. Educators and policy makers from the Tri-State area came to the booth to learn about protein structures, the RCSB PDB, and to take home tRNA tattoos.

• The RCSB PDB exhibited at the Biophysical Society's annual meeting (February 2-6; Long Beach, CA).

## Protein Sculptures on Display at Rutgers

*Cycloviolacin (2007, copper-coated steel, 24" x 32" x 36")*

Sculptures and photographs by Julian Voss-Andreae were on display at the Rutgers Student Center in New Brunswick, New Jersey in February.

Voss-Andreae's unique sculptures are designed to tell stories about hemoglobin, collagen, and other structures essential to life.

Julian Voss-Andreae is a German-born sculptor based in Portland. He graduated from the Pacific Northwest College of Art (PNCA) in 2004 with a BFA in sculpture. While still at PNCA, Voss-Andreae developed a novel kind of sculpture based on the structure of proteins, the building blocks of life. Voss-Andreae's work has been commissioned internationally and has been highlighted in journals such as *Leonardo* and *Science*.

Photographs of Voss-Andreae's sculptures are part of the RCSB PDB's Art of Science traveling exhibit, which also features images available from the RCSB PDB website and the *Molecule of the Month*. For more information on hosting this exhibit, please contact **info@rcsb.org**.

The next stop for the cycloviolacin sculpture is the *Art and Mathematics: The Wonders of Numbers* exhibit at The Heckscher Museum of Art, April 12 - June 22, 2008, in Huntington, NY.

Julian Voss-Andreae: **www.julianvossandreae.com**

The Heckscher Museum: **www.heckscher.org**

## Papers Published

The PDB archive, which began in 1971 as a handwritten petition signed by crystallographers, has developed into an online biological database and resource used by a diverse community of teachers, students, and researchers in academia and industry worldwide. This history is described in an article published in an issue of *Acta Crystallographica* that commemorates various milestones in the crystallographic community:

Helen M. Berman (2008) The Protein Data Bank: a historical perspective *Acta Cryst.* **A64**: 88-95. doi: 10.1107/S0108767307035623

A paper describing the work done as part of the wwPDB Remediation Project, including the standardization of IUPAC nomenclature for chemical components, an update of sequence database references and taxonomies, and improvements in the representations of viruses, has been published in *Nucleic Acids Research's 2008 Database Issue.*

K. Henrick, Z. Feng, W. F. Bluhm, D. Dimitropoulos, J. F. Doreleijers, S. Dutta, J. L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, C. L. Lawson, J. L. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, E. L. Ulrich, W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin, H. M. Berman (2008) Remediation of the Protein Data Bank archive *Nucleic Acids Research* **36**: D426-D433. doi: 10.1093/nar/gkm937

The data from the Remediation Project are available through the FTP archive and wwPDB member sites. Detailed documentation about the Remediation Project is available at **www.wwpdb.org**.

## Education Corner by Dr. Jeramia Ory, Kings College

### Moving Pictures: Using Chimera to make molecular multimedia for the classroom

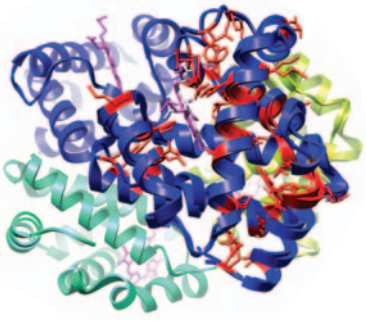
**JERAMIA ORY** *is an Assistant Professor in the Department of Biology at King's College in Wilkes-Barre, Pennsylvania where he teaches* Genetics, Biochemistry, *and* Systems Biology. *His training is in X-ray crystallography and NMR spectroscopy, and was previously a Biochemical Information Specialist at the RCSB PDB. While at the RCSB PDB, he produced numerous images for this newsletter, annual report and official documentation. The movies and figures he uses for class can be viewed on his homepage (**staff.kings.edu/jeramiaory**) in the "Multimedia" section, and are free for educational use.*

Getting students to grasp the link between 3D structure and biological function is a necessary and challenging part of many undergraduate courses. Structural information can help students "get it" in a way that cannot be underestimated. As an example, numerous students have told me how much easier it is to understand stereochemistry when they can manipulate chemicals on a computer screen in 3D rather than trying to work out wedge/hash 2D conventions. As the number of structures in the PDB archive continues to grow, the challenge lies not in finding structural information related to the topic at hand (the Advanced Search on the RCSB PDB website is a great resource), but in incorporating the information into lecture materials and presentations without draining an instructor's time or resources. Fortunately for instructors, a number of free programs that excel in molecular visualization and analysis are now available. Instead of reviewing the myriad of programs out there, I will focus the one I use to create multimedia presentations for my students–Chimera[1].

**CHIMERA** is written and maintained by the Computer Graphics Lab at the University of California, San Francisco. It has a long history in molecular visualization, having started as a program designed in 1980 for high-end graphics workstations. What this means practically is that this research group has been thinking about the needs of the molecular visualization community for a long time. As modern desktop computing power has grown, the visualization community has expanded from its original base of X-ray crystallographers to educators and students as young as high school. While no program can be all things to all people, Chimera comes close. I have personally used it for hands on molecular visualization workshops with groups ranging from high school students to undergraduates with good results. Chimera has a few advantages when compared to other packages out there.

**COST.** Chimera is free for academic use. This is becoming less of a unique feature with the rise of open source and free software, but it is still an important consideration. After using Chimera for an exercise, I direct students to the download page so they can use it on their home computers if they wish to continue exploring. Just as important, Chimera is available for every major operating system: Windows, Macintosh and Linux (and more). Of course, as critics of free software are fond of saying, "free software is only free if your time has no value." Luckily, the program is forgiving to new users, and rewards time spent with it.

**LEARNING CURVE.** The last thing educators have is time to waste. Chimera is a powerful analysis and visualization tool and is written with scientists in mind, however, it is quite easy to learn and new users can generally find their way around the program in about an hour. I run a protein visualization exercise in my undergraduate Biochemistry class that walks the students through the basics of Chimera; they align myoglobin and hemoglobin and then color the aligned residues by conservation (an example is shown). Most students complete the exercise in 50 minutes and find it useful to be able to explore protein structure on their own. Should they not finish, the fact that it is free means they can finish up at the campus computer lab or at home. The program is well-documented online (**www.cgl.ucsf.edu/chimera**) and comes with tutorials for new users.



*Myoglobin (2MM1) and hemoglobin (4HHB) aligned; residues are colored by conservation*

**SUPPORT.** The Chimera community continues to grow as the program reaches different user groups. There is an active mailing list of Chimera users that shares ideas and problems, and is an excellent resource. Furthermore, the developers of the program monitor the list, and have been known to write modules for the program to deal with special users requests. These requests have even been known to make it into future releases as new features. One way or another, you can usually find someone willing to help you make the figures or movies you want.

Pretty pictures, pretty fast. Let's face it, you can stand in front of a lecture hall for an hour, waving your arms, talking about the symmetrical relationships of hemoglobin's four subunits, or you can display some nicely ren-

dered images and a movie or two and get the same point across in five minutes. Say what you will about today's students, they respond to multimedia and in some cases have grown to expect it. This is where Chimera shines. Once learned, gorgeous images take just a few minutes to set up and render. In my biochemistry class, I often make structural figures rather than using the textbooks illustrations, or load Chimera in class and walk students through the structure in 3D. This accomplishes two things: 1) I get to highlight what it is I find important in an RNA double helix, an enzyme active site, *etc.*, and 2) it forces me to learn the structural landscape of the molecules rather than relying on the textbook. The rendering styles of Chimera are of course a matter of taste, but frequent readers of the RCSB PDB's newsletter have already seen what Chimera can do; it is the "go to" program among the staff and is used on many official publications.

So, now that you are convinced to give Chimera a try, how do we make some movies? There are tutorials on the web site, but to start with, you can try this simple set of commands. In your favorite text editor (Notepad in Windows or TextEdit in Mac OS X), create a file called "movie.cmd" and enter the following text:

```
movie record
roll y 1 360
wait 360
movie stop
movie encode
```

Save the file, start Chimera and load your molecule of interest. Get it looking how you like, then open the file "movie.cmd" (File… Open…) and watch it go. This script will rotate the molecule about the y axis in 360 steps of 1 degree, then save the movie. When it's done you should have a file named "chimera_movie.mov" that you can show to your students. If you would like to see some of the movies I use in my *Genetics* and *Biochemistry* lectures, or if you would like to use them in your own classes, please visit my website.

1. E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* **25**(13): 1605-12.

## PDB Community Focus

### Dr. Christine Orengo, University College London

**Q.** ***In 1997, you and your colleagues established CATH[1]–a system that is used to classify protein domain structures. How are researchers using CATH today? What types of research and discoveries does it enable? Has its usage changed in the past ten years?***

**A:** In the early 1990s, there were over three thousand structures deposited in the PDB and Janet Thornton realized that we could get some very useful insights into protein folding and evolution by grouping these into fold groups and evolutionary families. I was fortunate to join her group at that time and we set about doing this classification with the benefit of a very sensitive structure comparison algorithm developed by Willie Taylor and myself, at NIMR. We designed a hierarchical classification which grouped proteins according to their basic secondary structure composition (Class), 3D shape (Architecture), folding arrangement (Topology), and finally evolutionary ancestry (Homology). Although we largely use automated approaches, identifying domain boundaries in multi-domain proteins, and recognizing homologues are difficult and very time consuming, as they need manual validation, which is why we only have ~80% of the PDB classified to date. We have just introduced some sophisticated new protocols that we think will help us to increase this percentage over the next year.

Despite this slight lag with the PDB, CATH is widely used and currently receives about a million web page hits per month from sites all over the world. We have put considerable effort into the design of the resource, trying to present the information in an intuitive and easily accessible form, and I believe this is reflected in its high usage. SCOP[2], a related resource, is also very widely used but because we exploit slightly different criteria to classify folds and provide additional information on superfamilies (*e.g.* multiple structure alignments), the two resources are somewhat complimentary. I think CATH is particularly useful for teaching. Perhaps the other distinctive feature of CATH is that we have developed our own structure comparison methods and provide a service (CATHEDRAL web server)[3] for scanning new structures against representative domains. This is very popular with structural biologists as it can be used to recognize novel folds or classify new structures into existing superfamilies. The CATH fold library is also exploited by computational biologists developing methods to predict whether a sequence is likely to adopt one of the known structures.

We have now extended CATH to include all sequences in the genomes that can be predicted to belong to a CATH superfamily (CATH-Gene3D)[4] and this has allowed us to increase the functional annotations associated with each superfamily hugely. Biologists are increasingly using CATH and Gene3D to obtain structural and functional annotations for their proteins and this has been facilitated by further dissemination of the information through the DAS annotation systems set up by the Biosapiens network (**www.biosapiens.info**).

**CHRISTINE ORENGO** *is a Professor of Bioinformatics in the Structural and Molecular Biology Research Department of University College London (UCL). She studied chemical physics at Bristol University and was awarded a Ph.D. in enzyme kinetics at UCL in 1984. Following a brief spell in industry, she worked as a research fellow at the National Institute for Medical Research (NIMR) in London before moving back to UCL in 1992 to pursue further postdoctoral studies. She was awarded a Senior Research Fellowship by the Medical Research Council in 1995 and was appointed Chair in Bioinformatics in 2002. Together with Janet Thornton, she established the CATH domain structure classification in 1993 which led to the discovery of some highly populated fold groups in nature–the so-called superfolds.*

*Her current research interests are in structural, functional, and comparative genomics. Computational analyses exploit the CATH database of structural families and the more recently established sister resource for domain and protein families in completed genomes, Gene3D. She collaborates with a number of experimental groups involved in studying pain, cancer and host-viral interactions. She also participates in several European networks for genome annotation (Biosapiens), grid technologies (EMBRACE), and systems biology (ENFIN) and is a member of the NIH-funded PSI Midwest Center for Structural Genomics (MCSG) headed by Andrzej Joachimiak. She was one of the founding researchers of the bioinformatics based Inpharmatica company. She has authored over 150 papers, book chapters and reviews and is on the editorial board of* FEBS, BMC Structural Biology, PEDS, *and the* Journal of Structural and Functional Genomics. *She is on the advisory board of the Swiss Institute of Bioinformatics and the Marie Nostrum Supercomputer Centre in Barcelona.*

Perhaps one of the most interesting phenomena revealed by classifying structures is the incredible bias in the populations of the fold groups and evolutionary superfamilies. In 1994, Janet Thornton and I reported the existence of the superfolds, a set of 10 folds which were highly over-represented in CATH[5]. This trend still exists and the integration of sequence data through Gene3D has shown that it is not an artifact of sampling but a genuine reflection of the dominance of certain folds in nature. The bias is also apparent at the evolutionary superfamily level. For instance, the 100 largest superfamilies in CATH account for nearly half the domain sequences of predicted structures in completed genomes.

As CATH has become more highly populated, it has been used to study and characterize the structural mechanisms involved in the evolution of proteins and their functions; in particular, the extent to which structural embellishments to the domain core can modify the geometry of active sites or influence surface features mediating different protein-protein interactions. The integration of genome sequences in CATH-Gene3D has illuminated functional diversity across superfamilies, and recent changes in the usage of CATH reflects biologists' interests in performing comparative genome analyses with this extensive functional data. For example, a comparison of CATH superfamilies, universal to bacteria, revealed that the expansion of metabolic and regulatory superfamilies with genome size is balanced, allowing maximum enrichment of the metabolic repertoire within the constraints of maintaining a small genome for fast replication.[6]

**Q.** ***Do you think that we are close to having representatives of every possible fold? Have the structural genomics projects had an impact?***

**A.** I think this depends on one's definition of a fold. The huge structural diversity apparent in some of the largest CATH superfamilies has challenged my belief in a rigid hierarchical classification whereby relatives in each evolutionary superfamily adopt the same fold. For example, there is great structural diversity in many of the 100 most highly populated superfamilies, and there are clear examples of relatives with different folds. Whilst these relatives share 40-50% of residues in the cores of their structures, these cores can be embellished so differently that many structural biologists would say that the domains belong to different fold groups. That said, for the remaining ~2000 superfamilies, relatives can be characterized within a single fold group and so I feel that the topology or fold group level in CATH is still valuable.

The structural genomics initiatives, particularly the PSI initiative in the States which has the goal of solving novel folds and aims to determine structures for all large protein families, are helping both to increase the numbers of known folds in the PDB and also to address the question of whether the hundreds of thousands of apparently novel superfamilies in the genomes are truly novel, adopting folds that are distinct from anything seen before. These initiatives have been very successful in increasing the numbers of new folds deposited in the PDB each year. For example, over the last two years a large proportion of the novel folds in the PDB have come from the four major centers associated with this initiative. Interestingly, although PSI deliberately targets superfamilies thought to be unrelated to any known superfamilies in SCOP or CATH, only about 30% turn out to be new superfamilies with distinct folds once their structures are solved. The remainder have been found to be distant relatives of known fold groups and families.

As to whether we have representatives of every possible fold, our analysis of genome data using sensitive threading algorithms like David Jones's GenThreader[7] suggests that within each organism about 80% of sequences can now be assigned to one of ~1100 CATH folds. Thus I would say that we do have fold representatives for most of the major superfamilies in nature. However, nearly half of these predicted structures belong to the 100 very structurally-diverse superfamiles and so it is possible their folds may be slightly different to those already characterized.

Sequences which can't be assigned a fold in CATH tend to belong to very small superfamilies which are species-specific. The number of these superfamilies is growing enormously as the metagenomics initiatives continue. For example, sampling of bacterial proteins from different environments like the Sargasso sea, diverse soils and even the human gut, suggests the existence of hundreds of thousands of very small families and orphan sequences for which we have no structural data at present. Although some of these may be genuinely new superfamilies with folds never seen before, it is more likely that a significant proportion will be found to be distant relatives of structurally characterized families. Some divergence in the structure of these remote relatives would be likely as the different environmental contexts would probably result in the evolution of different functions, and this is frequently mediated by changes in the structure.

The problem in estimating the number of folds that remain to be determined lies in the currently rather subjective approaches used for defining fold similarity. If we assume two domains have similar folds–if they superpose with an RMSD less than 5Å, (normalized for the number of equivalent residues)–we actually find that there are nearly three times as many folds in CATH than represented by the ~1100 fold groups. Practically all of this increase is due to the structural diversity occurring across the 100 largest superfamilies. Since we know that a significant proportion of sequences from each organism are typically assigned to these very large superfamilies, as increasing numbers of structures are solved from different species, the total number of folds will grow, simply from an expansion of these very large superfamilies. In addition, since the thousands of new families arising from the metagenomics will either have novel folds or very likely be distant relatives of these very large superfamilies, and therefore with slightly different folds, over the next decade we could certainly see hundreds more structures which are rather different from any known folds, especially if the structural genomics initiatives continue to be funded.

However, as I mentioned before, whether we view these as completely new folds depends on our definition of fold. We no longer refer to the T level in CATH as the topology or fold group level but rather the 'topological motif' or 'fold motif' level. In other words, structures grouped at this level share a large central structural motif or core 'fold motif' comprising about 40-50% of the domain's residues. I believe that the majority of 'fold motifs' in nature have now been characterized, with the structural genomics contributing significantly to this repertoire of fold motifs over the last decade. Structures remaining to be solved are highly likely to have core motifs similar to one of the ~1100 fold motifs characterized in CATH or SCOP, but these cores may be structurally decorated in ways not seen yet. By improving the characterization of these fold motifs and understanding the manner in which they can be structurally embellished, we hope to improve the structural annotation and modeling of all the sequence relatives in the genomes.

Regardless of the definition of a fold, we are interested in discovering all the different ways in which proteins fold into their 3D dimensional structure and interact with ligands. Gene3D was established to structurally annotate the genomes and integrate functional data from all the sequences. Using these data, we can better understand the structure-function relationship with respect to protein-protein and protein-ligand interactions. Although the ways in which proteins bind ATP could be limitless, they are likely to be very similar in proteins with the same fold. Therefore, by targeting predicted new folds and diverse functional subfamilies the structural genomics initiatives should deepen our understanding of protein folding and protein-ligand interactions and move us further towards a structure-function model for all proteins.

**Q.** ***Predicting the function of a given protein is a great challenge. How do Gene3d and the PDB archive play into this type of research?***

**A.** Clearly the value of structures solved by the structural genomics ini-

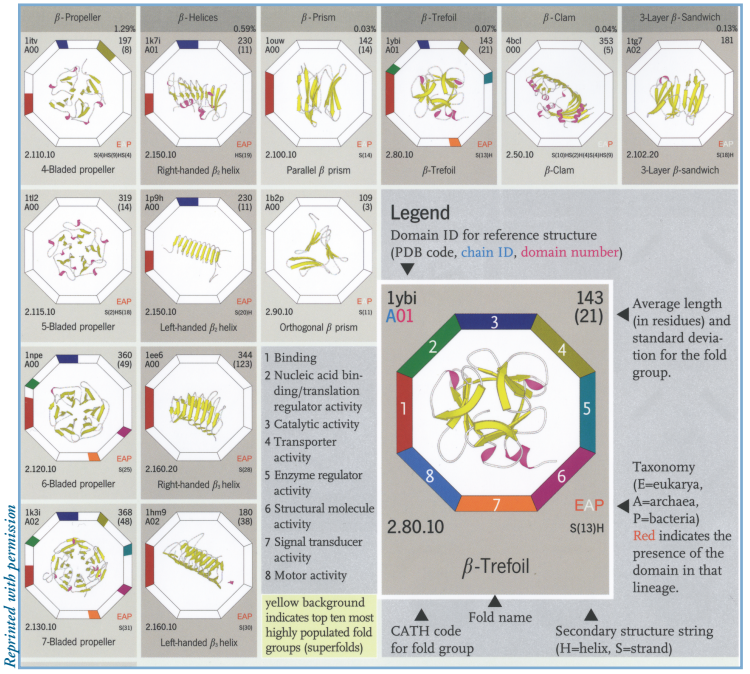
tiatives increases once functions become known for them or as methods for predicting function from structure improve. To facilitate this, we have been increasing the amount of functional information stored in Gene3D. Fortunately, there are now many excellent public resources providing functional information. Those captured in Gene3D include GO[8], COGs,[9] FunCat,[10] and EC[11] amongst others. We carefully inherit this functional information between relatives using various bioinformatics protocols. Some approaches exploit simple pair-wise sequence identity between relatives whilst others use more sophisticated methods (*e.g.* HMM-HMM comparisons) to allow safe inheritance between more distant relatives sharing common functions. Knowledge of the pathways or biological processes that a protein participates in is also useful for understanding its functional role, and so we have incorporated information on protein interactions in Gene3D (*e.g.* from KEGG,[12] Reactome,[13] and IntAct[14]) and developed a suite of bioinformatics tools for predicting interactions between proteins, too. Integrating data in Gene3D in this way allows us to draw together as much collated information on genes as possible both to enhance biomedical research, as well as our model of protein evolution.

The recently created PSI Structural Genomics Knowledgebase (**kb.psi-structuralgenomics.org/KB**) will help enormously in extending the functional information available for each structure. With the aim of integrating and presenting functional information from a wide range of public resources, this will significantly enhance structural studies on how proteins function. In addition, other initiatives such as the EU-funded Biosapiens network for structural and functional annotation of genomes will also play a part in providing functional annotations for PDB structures. A recent analysis performed for the Midwest Center for Structural Genomics showed that by using Gene3D, some functional information could be gleaned for a large proportion of sequences targeted for structure determination. Some of this is rather general information and may not be that useful at present, except in directing further experiments (*e.g.* mutation experiments) but a reasonable proportion is detailed enough to allow some mechanistic rationale to be derived from the solved structure.

Furthermore, since recent aims of the PSI structural genomics initiatives include targeting additional relatives from the most highly populated CATH superfamilies, relatives can be targeted which are predicted to be functionally diverse from those with close homologues of known structures. Expanding the repertoire of structures for different functional subfamilies within these superfamilies will increase our understanding of structure-function relationships and ultimately improve function prediction methods. Recent analyses of structures of unknown function solved by the Midwest consortium using the ProFunc resource developed by the Thornton group, showed that some functional information could be predicted for a large proportion of the structures. This success rate is likely to increase as structural genomics initiatives deliberately target sequences with known functions and the resulting increase in coverage of structure-function space improves our function prediction algorithms.

**Q.** ***With Richard C. Garratt, you've recently published a great educational tool called* The Protein Chart.[15] *What was the inspiration for this "periodic table" of proteins? How do you think it will be used?***

**A.** Richard and I really enjoyed developing this chart and we had two excellent CATH researchers in my group, Alison Cuff and Ian Sillitoe, who made the whole project possible. The idea for a protein chart originally arose from the structure modeling kit that Richard had designed for Wiley which is a wonderful teaching tool for explaining how structures are built from their component secondary structures. It's really a Lego toolkit for proteins! Wiley wanted a protein chart showing examples of representative structures that students could try to build with the kit. We were very excited by the project and inspired to produce a design based on the ideal 'periodic chart' of protein structures proposed by Willie Taylor a few years ago.[16] This shows simple representations of all the types of architectures or




*A section of the* β*-proteins shown in* The Protein Chart *(www.wiley.com)*

3D protein shapes that should be seen in nature given the rules drawn up over the last three decades for protein folding and packing. We thought Willie's chart was a wonderful way of representing our current knowledge of protein architectures and imagining what shapes and folds remained to be discovered.

So we designed a protein chart, arranged like a periodic table, but showing representatives of all the domain architectures or shapes currently deposited in the PDB and classified in CATH. There are over 30 different architectures in CATH which are regular enough for the 2D image of the structure to provide meaningful information, and for each of these, the chart shows the ranges of sizes observed. The chart also contains information on the proportion of genome sequences that are predicted to adopt each type of shape, and also the types of functions exhibited in the different fold groups. There are also illustrations of common supersecondary motifs and oligomeric proteins, and so we think it will be a very useful tool for undergraduate teaching and also for structural biology researchers. I would imagine that computational biologists developing structure prediction methods will also find it a useful way of learning about the different shapes and folds they are trying to predict. With the progress of the structural genomics initiatives, especially the PSI, in solving novel structures, we can expect the chart to evolve and expand over the next decade and it will be a useful visual aid for monitoring our knowledge of the structural universe.


[1] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton (1997) CATH–a hierarchic classification of protein domain structures. *Structure.* **5**: 1093-1108.
[2] L. Conte, A. Bart, T. Hubbard, S. Brenner, A. Murzin, C. Chothia (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **28**(1): 257-259.
[3] O.C. Redfern, A. Harrison, T. Dallman, F.M. Pearl, C.A. Orengo (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol.* **3**(11): e232.
[4] C. Yeats, J. Lees, A. Reid, P. Kellam, N. Martin, X. Liu, C. Orengo (2008) Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res.* **36**: D414-8.
[5] J. Ranea, D. Buchan, J. Thornton, C. Orengo (2005) Microeconomic principles explain an optimal genome size in bacteria *Genetics.* **21**: 21-25.
[6] C.A. Orengo, D.T. Jones, J.M. Thornton (1994) Protein superfamilies and domain superfolds. *Nature.* **372**(6507): 631-4.
[7] D.T. Jones (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol.* **287**(4): 797-815.
[8] The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics.* **25**: 25-29.
[9] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**: D13-21.
[10] A. Ruepp, A. Zollner, D. Maier, K. Albermann, *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* **32**(18): 5539-45.
[11] Enzyme Nomenclature, Enzyme Classification. **www.chem.qmw.ac.uk/iubmb/enzyme.**
[12] M. Kanehisa and S. Goto (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1): 27-30.
[13] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**(Database issue): D428-32.
[14] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.* **35**(Database issue): D561-5.
[15] R.C. Garratt and C. Orengo (2008) *The Protein Chart* Weinheim: Wiley-VCH.
[16] W.R. Taylor (2002) A 'periodic table' for protein structures. *Nature* **416**: 657-60.

## RCSB PDB Partners

The RCSB PDB is managed by two partner sites of the Research Collaboratory for Structural Bioinformatics:

RUTGERS

Rutgers, The State University of New Jersey
Department of Chemistry and
Chemical Biology
610 Taylor Road
Piscataway, NJ 08854-8087

UCSD SDSC • SKAGGS SCHOOL of PHARMACY and PHARMACEUTICAL SCIENCES

San Diego Supercomputer Center and the Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0537

WORLDWIDE PDB PROTEIN DATA BANK

The RCSB PDB is a member of the Worldwide Protein Data Bank (www.wwpdb.org)

## RCSB PDB Management

**DR. HELEN M. BERMAN**, Director
Rutgers, The State University of New Jersey
berman@rcsb.rutgers.edu

**DR. PHILIP E. BOURNE**, Associate Director
San Diego Supercomputer Center and the Skaggs School of Pharmacy and Pharmaceutical Sciences,
University of California, San Diego
bourne@sdsc.edu

**DR. MARTHA QUESADA**, Deputy Director
Rutgers, The State University of New Jersey
mquesada@rcsb.rutgers.edu

A list of current RCSB PDB Team Members is available from www.pdb.org.

**STATEMENT OF SUPPORT:** *The RCSB PDB is supported by funds from the National Science Foundation, the National Institute of General Medical Sciences, the Office of Science, Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering, the National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes & Digestive & Kidney Diseases.*

**RCSB PROTEIN DATA BANK**
www.pdb.org

Department of Chemistry and Chemical Biology
Rutgers, The State University of New Jersey
610 Taylor Road
Piscataway, NJ 08854-8087
USA

*Return Service Requested*