RCSB PDB NEWSLETTER

NUMBER 20 ◆ WINTER 2004 ◆ *Published quarterly by the Research Collaboratory for Structural Bioinformatics Protein Data Bank*

Weekly RCSB PDB news is available on the Web at www.rcsb.org/pdb/latest_news.html

Links to RCSB PDB newsletters are available at www.rcsb.org/pdb/newsletter.html

CONTENTS

SNAPSHOT: JANUARY 1, 2004

23,792 released atomic coordinate entries

MOLECULE TYPE		EXPERIMENTAL TECHNIQUE	
21,516	proteins, peptides, and viruses	20,224	diffraction and other
1,276	nucleic acids	11,469	structure factor files
982	protein/nucleic acid complexes	3,568	NMR
18	carbohydrates	1,790	NMR restraint files

PARTICIPATING RCSB MEMBERS

RUTGERS: rutgers.rcsb.org

SDSC/UCSD: www.pdb.org

CARB/NIST: nist.rcsb.org

E-MAIL: info@rcsb.org

FTP: ftp.rcsb.org

MESSAGE FROM THE RCSB PDB

The RCSB PDB is excited to start the new year as part of a special collaboration with the Macromolecular Structure Database at the EMBL-European Bioinformatics Institute (MSD-EBI), and Protein Data Bank Japan (PDBj)—the Worldwide Protein Data Bank (wwPDB; www.wwpdb.org). The wwPDB is committed to a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community. All three organizations serve as deposition, data processing and distribution sites of this PDB Archive. Each wwPDB site provides its own view of the primary data, thus providing a variety of tools and resources for the global community. Details of this historic agreement are described in this newsletter. The formation of the wwPDB will be transparent to users and will ensure the overall quality and consistency of data directly available through the PDB.

In early 2004, RCSB PDB members have presented at the Pacific Symposium on Biocomputing (January 6–10, Hawaii) and will be exhibiting at the Biophysical Society Meeting (February 14–18, Baltimore, MD).

We wish all of our users a very happy new year!

The RCSB PDB ✦

The Protein Data Bank (PDB) is the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data. The PDB is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center (SDSC) at the University of California, San Diego (UCSD); and the Center for Advanced Research in Biotechnology (CARB) of the National Institute of Standards and Technology (NIST)—three members of the Research Collaboratory for Structural Bioinformatics (RCSB), a non-profit consortium dedicated to improving our understanding of biological systems.

MIRROR SITES

Cambridge Crystallographic Data Centre (UK): pdb.ccdc.cam.ac.uk

National University of Singapore: pdb.bic.nus.edu.sg

Osaka University (Japan): pdb.protein.osaka-u.ac.jp

Universidade Federal de Minas Gerais (Brazil): www.pdb.ufmg.br

Max Delbrück Center for Molecular Medicine (Germany): www.pdb.mdc-berlin.de

Announcing the Worldwide PDB

Reprinted with permission from **Nature Structural Biology.**

In recognition of the growing international and interdisciplinary nature of structural biology, three organizations have formed a collaboration to oversee the newly formed worldwide Protein Data Bank (wwPDB; **www.wwpdb.org**). The Research Collaboratory for Structural Bioinformatics (RCSB), the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (EBI) and the Protein Data Bank Japan (PDBj) at the Institute for Protein Research in Osaka University will serve as custodians of the wwPDB, with the goal of maintaining a single archive of macromolecular structural data that is freely and publicly available to the global community.

The wwPDB represents a milestone in the evolution of the Protein Data Bank (PDB; **www.pdb.org**[1,2], which was established in 1971 at Brookhaven National Laboratory as the sole international repository for three-dimensional structure data of biological macromolecules. Since July 1, 1999, the PDB has been managed by three member institutions of the RCSB: Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology.

The wwPDB recognizes the importance of providing equal access to the database—both in terms of depositing and retrieving data—from different regions of the world. Therefore, the wwPDB members will continue to serve as deposition, data processing, and distribution sites. Deposition procedures will not be altered by the formation of the wwPDB; data can still be deposited using ADIT at the RCSB and PDBj or by using AutoDep at the EBI.

To ensure the consistency of PDB data, all entries will be validated and annotated following a common set of criteria. All processed data will be sent to the RCSB, which distributes the data worldwide. All format documentation will be kept publicly available and the distribution sites will mirror the PDB archive using identical contents and subdirectory structure. However, each member of the wwPDB will be able to develop its own Web site, with a unique view of the primary data, providing a variety of tools and resources for the global community.

An Advisory Board consisting of appointees from the wwPDB, the International Union of Crystallography and the International Council on Magnetic Resonance in Biological Systems will provide guidance through annual meetings with the wwPDB consortium. This board is responsible for reviewing and determining policy as well as providing a forum for resolving issues related to the wwPDB. Specific details about the Advisory Board can be found in the wwPDB charter, available on the wwPDB Web site.

The RCSB is the 'archive keeper' of wwPDB. It has sole write access to the PDB archive and control over directory structure and contents, as well as responsibility for distributing new PDB identifiers to all deposition sites. The PDB archive is a collection of flat files in the legacy PDB file format[3] and in the mmCIF[4] format that follows the PDB exchange dictionary (**deposit.pdb.org/mmcif**). This dictionary describes the syntax and semantics of PDB data that are processed and exchanged during the process of data annotation. It was designed to provide consistency in data produced in structure laboratories, processed by the wwPDB members and used in bioinformatics applications. The PDB archive does not include the Web sites, browsers, software and database query engines developed by researchers worldwide.

The members of the wwPDB will jointly agree to any modifications or extensions to the PDB exchange dictionary. As data technology progresses, other data formats (such as XML) and delivery methods may be included in the official PDB archive if all the wwPDB members concur on the alteration. Any new formats will follow the naming and description conventions of the PDB exchange dictionary. In addition, the legacy PDB format would not be modified unless there is a compelling reason for a change. Should such a situation occur, all three wwPDB members would have to agree on the changes and give the structural biology community 90 days advance notice.

The creation of the wwPDB formalizes the international character of the PDB and ensures that the archive remains single and uniform. It provides a mechanism to ensure consistent data for software developers and users worldwide. We hope that this will encourage individual creativity in developing tools for presenting structural data, which could benefit the scientific research community in general. ◆

W O R L D W I D E
wwPDB
P R O T E I N   D A T A   B A N K

[1] *H.M. Berman, et al. (2000):* Nucleic Acids Res. *28, pp. 235-242.*

[2] *F.C. Bernstein, et al. (1977):* J. Mol. Biol. *112, pp. 535-542 .*

[3] *J. Callaway, et al. (1996): Protein Data Bank Contents Guide: Atomic coordinate entry format description. (Brookhaven National Laboratory).*

[4] *P.E. Bourne, H.M. Berman, K. Watenpaugh, J.D. Westbrook, & P.M.D. Fitzgerald (1997):* Methods Enzymol. *277, pp. 571-590.*

DATA DEPOSITION AND PROCESSING

Downloadable PDB_EXTRACT Makes Deposition Easier

The software program PDB_EXTRACT was developed to assist depositors in the automatic preparation of crystallographic depositions. This software tool extracts information for deposition from the output files produced by many applications used for structure determination. Current versions of the following programs are supported:

HKL 2000	SHARP	PHASES	CNS
SCALEPACK	autoSHARP	SnB	REFMAC
d*TREK	SHELXD	BnP	RESTRAIN
SOLVE	SHELXE	DM	TNT
MLPHARE	SHELXL	Solomon	WARP
		RESOLVE	

PDB_EXTRACT will also be part of the CCP4 Program Suite (version 5).

Files produced by PDB_EXTRACT can be edited on a local Linux workstation using the downloadable version of ADIT, which has been extended to provide access to the large amount of data collected by the PDB_EXTRACT program.

PDB_EXTRACT can be downloaded in source and binary versions for Linux, SGI, SUN, OSF and Mac OSX from deposit.pdb.org/software. Source and Linux binary versions of ADIT are also available.

Questions and comments may be sent to help@rcsb.rutgers.edu.
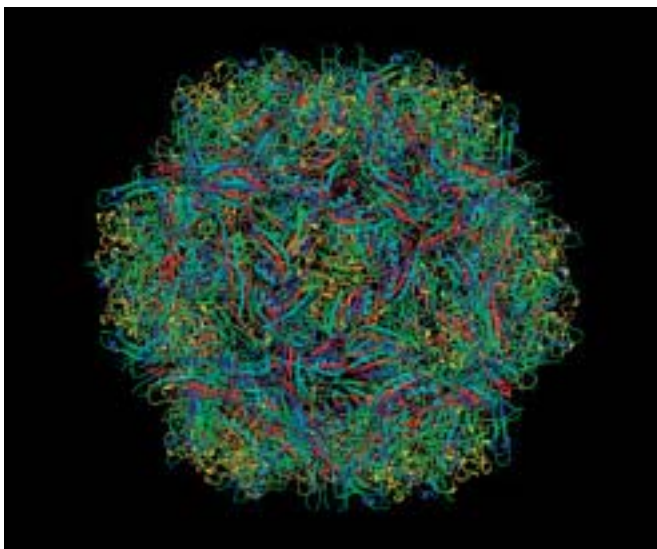
In the structure of the host range-controlling region of feline parvovirus in entry 1p5y, non-crystallographic symmetry is used to create the icosohedral viral capsid from sixty copies of the one protein chain contained in the coordinate file. The viral capsid is both the asymmetric unit and the biological unit.

PDB ID: **1p5y**

L. Govindasamy, K. Hueffer, C.R. Parrish, M. Agbandje-McKenna (2003): Structures of host range-controlling regions of the capsids of canine and feline parvoviruses and mutants. J. Virol. 77, p. 12211.

Biological Unit Tutorial Now Available from the RCSB PDB

An introduction to biological units in the PDB archive is now accessible at www.rcsb.org/pdb/biounit_tutorial.html. This useful guide offers detailed explanations and examples of the asymmetric unit and the biological molecule, indicates where information about the biological unit can be found in PDB and mmCIF coordinate files, and describes how the biological unit files in the PDB have been derived.

The RCSB PDB offers images and coordinates for the complete biological unit of crystallographic entries in the archive. The biological molecule, or biological unit, is the macromolecule that has been shown to be or is believed to be functional. When crystallographic structures are deposited into the PDB, the primary coordinate file generally contains one asymmetric unit—the smallest portion of a crystal structure to which crystallographic symmetry can be applied to generate one unit cell. In some cases, the asymmetric unit differs from the biologically active molecule. The RCSB PDB provides data on biological units to further an understanding of molecular function.

The biological unit tutorial is also linked from the View Structure and Download/Display File sections of the Structure Explorer page, as well as under PDB WWW User Guides. For more information, please send inquiries to info@rcsb.org.

Ligand Depot—a Small Molecule Information Resource

Ligand Depot (ligand-depot.rutgers.edu) is a data warehouse that integrates databases, services, and tools related to small molecules bound to macromolecules. The initial release (v. 1.0, November, 2003) focuses on providing chemical and structural information for ligands that are found as part of the structures deposited with the PDB.

Ligand Depot allows users to extract ligand information from the PDB, to perform chemical substructure searches, and to search other small molecule resources on the Web. One of the distinguishing features of Ligand Depot is that it allows users to retrieve the coordinates of any small molecule found within the structure entries of the PDB. It is also updated daily and therefore provides the most current information on small molecules present in the PDB.

Ligand Depot currently includes chemical descriptions for the ~4,600 ligands that are part of the structures deposited in the PDB, and it offers various search options for obtaining information on these small molecules. It accepts keyword queries based on PDB ligand code, compound name and chemical formula. Using a simple graphical interface, a substructure search may also be performed between a small molecule of interest and all of the ligands present in the PDB.

Ligand Depot can also be used to browse a variety of other small molecule resources on the Web. Information from 70 small molecule sites are stored in Ligand Depot. These resources are organized into four categories, including molecular visualization sites, commercial sites, nomenclature sites, and chemical databases. Keyword searches may be performed on these external Web sites

if they are search-enabled. Thus, information on ligands can be extracted from a diverse collection of Web resources using a single query.

A helpful tutorial for using Ligand Depot is accessible at ligand-depot.rutgers.edu/html1/User_Guides.html.

The Ligand Depot interface

PDB Focus: Deposition and Release Policies

G uidelines for the deposition of coordinate and experimental data have been set by the IUCr, IUPAC-IUBMB-IUPAB, the NIH, and the journals. These policies are detailed at deposit.pdb.org/#release.

Depending upon the hold status selected by the depositor, data release occurs when a depositor gives approval (REL), the hold date has expired (HOLD), or the journal article has been published (HPUB).

As of May 6, 2003 (www.rcsb.org/pdb/pdb_news2003.html#hpub), there is a one-year limit on the length of a hold period, including HPUBs. If the citation for a structure is not published within the one-year period, depositors will be given the option to either release or withdraw the deposition.

Detailed deposition and release information is available at deposit.pdb.org/#release.

PDB Deposition Statistics

I n 2003, 4,831 structures were deposited to the PDB archive, and were processed by teams at RCSB-Rutgers, Osaka University, and the European Bioinformatics Institute.

Of the structures deposited, 78% were deposited with a release status of "hold until publication"; 14% were released as soon as annotation of the entry was complete; and 8% were held until a particular date.

83% of these entries were determined by X-ray crystallographic methods; 13% were determined by NMR methods.

57% of these depositions released the sequence in advance of the structure's release. 72% of these depositions were deposited with experimental data.

Data Query, Reporting, and Access

Lucene Keyword Search Released on the RCSB PDB Web Site

A fter a period of beta testing, the Lucene keyword search engine has replaced the previously-used LDAP keyword search engine to support text searches on the RCSB PDB home page, SearchLite, and the "Text Search" field on SearchFields. Lucene uses an index of the remediated mmCIF files to return much more accurate keyword search results.

Lucene supports wildcard searches, phrases, Boolean queries, and offers a spell checker. Options are offered to narrow the scope of the query, for example, to search for author names or PDB IDs; the default is set to search the entire text of the mmCIF file indices. Additionally, partial word and exact word matches are supported; the default is set to perform an exact word match, unless the partial word match option is selected. The home page keyword search will locate exact word matches to a query.

Examples of supported queries can be found on the SearchLite page at www.rcsb.org/pdb/searchlite.html, and additional help can be found at www.rcsb.org/pdb/help-searchlite.html.

PDB Focus: Redundancy Reduction Cluster Data Available on the PDB FTP Site

T he results of the weekly clustering of protein chains in the PDB are posted at ftp://ftp.rcsb.org/pub/pdb/derived_data/NR/. These clusters are used in the "remove similar sequences" feature on SearchLite, SearchFields, and the home page on the RCSB PDB Web sites.

Files that list the clusters and their rankings at 50%, 70% and 90% sequence identity are available. Smaller rank numbers indicate higher (better) ranking. Chains with rank number one are ranked as the best representative of their cluster.

The contents of these files and the details of the clustering and ranking are further described at ftp://ftp.rcsb.org/pub/pdb/derived_data/NR/README and www.rcsb.org/pdb/redundancy.html.

PDB Focus: Searching for Experimental Data Files

T he PDB offers several ways to locate experimental data files for structure entries. The SearchFields interface offers an option to narrow a search to only include entries that have experimental data (X-ray structure factors or NMR restraints) available. This option can be activated by selecting "Experimental Data Availability" from the custom options at the bottom of the SearchFields page. To further narrow the search to only structure factors or only constraint data, select the preferred experimental method from the pull down menu in the "Exp. Technique" field. Information about this and other options available on the Search-Fields interface can be found on the SearchFields help page.

Experimental data files can be downloaded from the Structure Explorer page, if the experimental data file is available for that structure. Click on the "Structure Factors" or "NMR Restraints"

link on the left side of the page to access the experimental data file.

Experimental data files are also available for downloading from the RCSB PDB FTP site. Directories for either X-ray structure factors or NMR restraints can be found at ftp://ftp.rcsb.org/pub/pdb/data/structures/all/, or subdivided by the second and third character of their PDB IDs at ftp://ftp.rcsb.org/pub/pdb/data/structures/divided/. Experimental data files for structures that have been removed from the archive can be found at ftp://ftp.rcsb.org/pub/pdb/data/structures/obsolete/.

A search for PDB entries that include experimental data can be performed from the SearchFields interface by selecting the "Yes" radio button in the Exp. Data Available field, as indicated in the screenshot above.

Updates of mmCIF Files on the RCSB PDB FTP Site

As previously announced, the update of September 2, 2003 included the replacement of all mmCIF files in the RCSB PDB FTP archives with the remediated mmCIF files. Due to our ongoing data curation efforts, occasional weekly updates will include the replacement of large numbers of mmCIF files. We have decided to reserve the first Tuesday of each month for these potential bulk mmCIF updates. Such updates should not be required every month. If you would like to be added to a list of FTP users who will receive individual e-mail notifications prior to each bulk update, please send your request to info@rcsb.org.

RCSB PDB Web Site Statistics

The PDB is available from several Web and FTP sites located around the world. Users are also invited to preview new features at the RCSB PDB beta test site, accessible at beta.rcsb.org/pdb.

The access statistics are given below for the primary RCSB PDB Web site at www.pdb.org. ✦

Access Statistics for www.pdb.org

MONTH	DAILY AVERAGE			MONTHLY TOTALS		
	HITS	FILES	SITES	KBYTES	FILES	HITS
Dec 03	205,861	156,229	105,860	386,567,809	4,686,883	6,175,837
Nov 03	242,988	183,154	115,022	341,406,392	5,311,492	7,046,665
Oct 03	249,774	192,435	133,618	485,948,515	5,773,072	7,493,230

RCSB PDB OUTREACH

NIGMS News: PSI-2 and Structural Biology Roadmap RFA

Concept Clearance of the PSI-2 Production Phase

Plans for the next phase of the NIGMS Protein Structure Initiative (PSI) were announced at the recent NIGMS Council meeting (www.nigms.nih.gov/news/reports/council-psi-sept03.html). This phase will begin in 2005 with the grant announcement expected for early 2004. It is envisioned as an interacting network with large-scale research centers that will operate as high throughput structural genomics pipelines for protein production and structure determination. The plans approved by the Council also include the establishment of specialized research centers for development of new methods, technology, and approaches for the production and structure determination of especially challenging proteins, such as membrane proteins and proteins from humans and other higher eukaryotic organisms, as well as for projects to address technology barriers to high-throughput operation.

Since 2000, the NIGMS has funded nine pilot structural genomics research centers as part of its plan to reduce the costs and increase the success of the structural determination of proteins. The long-range goal of the PSI is to make the three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences. The pilot projects have focused on high throughput methods for structure determination in order to achieve these goals. For more information please visit www.nigms.nih.gov/psi.

Announcement of Structural Biology Roadmap RFA

Structural biology is also prominent in the plans of the NIH Roadmap for Medical Research (nihroadmap.nih.gov/structuralbiology/index.asp). The roadmap includes an RFA (request for applications) for Centers for Innovation in Membrane Protein Production. Letters of intent are due by February 5, 2004 with applications due by March 11, 2004.

RCSB PDB Article Published in Nucleic Acids Research

The article, "The distribution and query systems of the RCSB Protein Data Bank," has been published in the latest issue of *Nucleic Acids Research*. This feature describes the dissemi-

nation and accessibility of PDB data via the current RCSB PDB query and distribution system. It also introduces an alpha version of the future re-engineered system that will be released in beta during the first quarter of 2004. The abstract and full text of the article are available from the *Nucleic Acids Research* Web site at nar.oupjournals.org.

P.E. Bourne, K.J. Addess, W.F. Bluhm, L. Chen, N. Deshpande, Z. Feng, W. Fleri, R. Green, J.C. Merino-Ott, W. Townsend-Merino, H. Weissig, J. Westbrook and H.M. Berman (2004): The distribution and query systems of the RCSB Protein Data Bank. Nucleic Acids Research *32, pp. D223-5.*

New Update Release of CD-ROM Sets

T he October 2003 update of the PDB CD-ROM data set, Release #106, is an incremental set of 1,583 experimentally determined structures and 61 models. The structure coordinate files, contained on one disk, are shipping now.

Files for entries re-released for any reason between July and October 2003 are included in this update. A list of files that have become obsolete since the last update is included so users can update their entire set of structures.

The first release of every year, in January, will include all structures. April, July and October updates will only contain the structures released during the previous quarter. New subscribers will receive the January release of the current year and all subsequent updates while supplies last.

The index files in the pub/resource subdirectory continue to include all structures in the current PDB FTP site as of that release.

Experimental data files—NMR constraints and X-ray structure factors—are released on the same schedule as the structure files: a complete set in January, and incremental updates for the three subsequent quarters.

Note: We are out of stock of Release #103, January 2003. New subscribers will be added to the list for Release #107, January 2004.

Questions should be directed to info@rcsb.org. Ordering information is available at www.rcsb.org/pdb/cdrom.html.

PDB Molecules of the Quarter: Trypsin, Simian Virus 40, and Catabolite Activator Protein

The *Molecule of the Month* series, by David S. Goodsell, explores the functions and significance of selected biological macromolecules for a general audience. These installments are available at www.rcsb.org/pdb/molecules/molecule_list.html. A sample of the molecules featured during this past quarter are included below:

Trypsin: Protein Cutting Machinery

October, 2003—Your body needs a steady supply of amino acids for use in growth and repairs. Each day, a typical adult needs something in the range of 35-90 grams of protein, depending on their weight. Quite surprisingly, a large fraction of this may come from inside. A typical North American diet may contain 70-100 grams of protein each day. But your body also secretes 20-30 grams of digestive proteins, which are themselves digested when they finish their duties. Dead intestinal cells and proteins leaking out of blood vessels are also digested and reabsorbed as amino acids, showing that our bodies are experts at recycling.
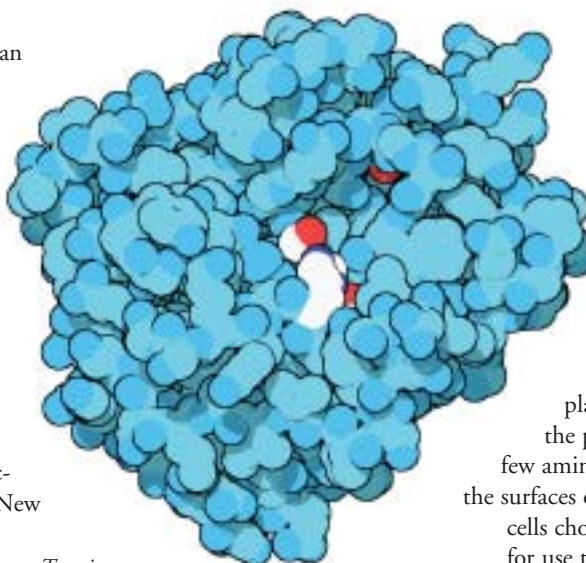
Trypsin
PDB ID: **2ptn**

J. Walter, W. Steigemann, T.P. Singh, H. Bartunik, W. Bode, R. Huber (1982): On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography. Acta Crystallogr.*, Sect. B 38, p. 1462.*

Proteins are tough, so we use an arsenal of enzymes to digest them into their component amino acids. Digestion of proteins begins in the stomach, where hydrochloric acid unfolds proteins and the enzyme pepsin begins a rough disassembly. The real work then starts in the intestines. The pancreas adds a collection of protein-cutting enzymes, with trypsin playing the central role, that chop the protein chains into pieces just a few amino acids long. Then, enzymes on the surfaces of intestinal cells and inside the cells chop them into amino acids, ready for use throughout the body.

Trypsin uses a special serine amino acid in its protein-cutting reaction, and is consequently known as a serine protease. The serine proteases are a diverse family of enzymes, all of which use similar enzymatic machinery. In digestion, trypsin, chymotrypsin and elastase work together to chop up proteins. Each has a particular taste for protein chains: trypsin (shown in PDB entry **2ptn**) cuts next to lysine and arginine, chymotrypsin (shown in PDB entry **2cha**) cuts next to phenylalanine and other large amino acids, and elastase likes chains with small amino acids like alanine (shown in PDB entry **3est**). Trypsin-like enzymes are also found in many other places in the body. Some of these are highly specific, cleaving only a specific target protein. For instance, thrombin, presented in the *Molecule of the Month* in January 2002, is designed to make a specific cut in fibrinogen, creating a blood clot.

For more information about trypsin, see www.rcsb.org/pdb/molecules/pdb46_2.html.

Simian Virus 40: Steering the Cycle of Life

November, 2003—Simian virus 40 is an example of how simple a virus can be and still perform its deadly job. Viruses are tiny machines with a single purpose: to reproduce themselves. They enter cells and hijack their synthetic machinery, forcing them to
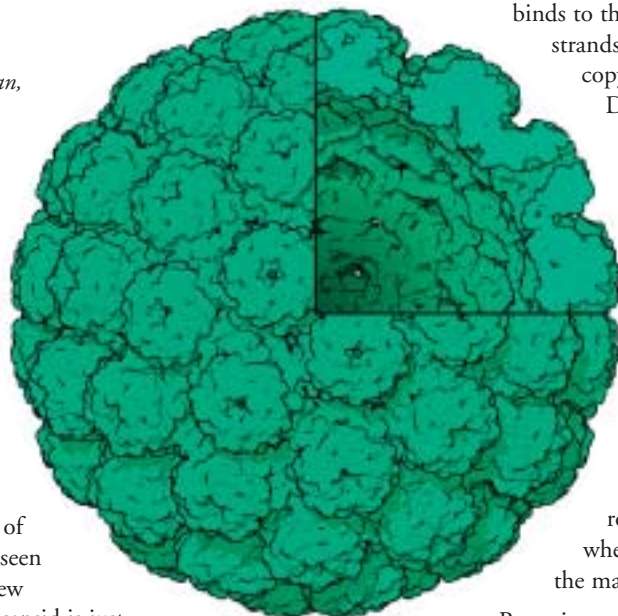
create new viruses. SV40 does this with very little molecular machinery. It is enclosed by a spherical capsid composed of 360 copies of one protein, seen in PDB entry **1sva**, and a few copies of two others. This capsid is just big enough to enclose a small circle of DNA 5,243 nucleotides long, which contains the barest minimum of information needed to get into the cell and make new viruses.

The circular SV40 genome is found in the cell as a "mini-chromosome" wound into a handful of nucleosomes, as shown here. It only has enough space to encode a few functions, since it all has to fit inside the tiny capsid. It has a regulatory region that controls the entire lifecycle of the virus. It also encodes several proteins: the T-antigen (and a spliced version of it called the t-antigen) and three capsid proteins, VP1, VP2 and VP3. Only a few tiny segments are not used. Space is so limited in this genome that the capsid proteins are actually encoded with overlapping reading frames, such that the end portion of the gene for one protein also encodes for the beginning portion of the next protein. For more information on the parsimonious genome of SV40, take a look at the European Bioinformatics Institute's *Protein of the Month* feature at www.ebi.ac.uk/interpro/potm/archive.html.

SV40 infects primate cells, forcing its way inside and releasing its DNA circle. Once inside, it has two jobs: to replicate its DNA and to package it inside new viral capsids. Amazingly, SV40 only needs one protein, the T-antigen, to control both of these processes. Soon after the virus enters the cell, the cell's own synthetic machinery recognizes a TATA sequence at the center of the SV40 regulatory regions. The cell then creates a messenger RNA reading counter-clockwise around the DNA circle. This mRNA is used to make the T-antigen protein. Then the virus really gets to work. The T-antigen

binds to the SV40 circle and helps to separate the strands, making way for the cell's polymerases to copy the DNA. It also directs the reading of the DNA in the opposite direction, clockwise around the strand, to create many copies of the capsid proteins.

For more information on simian virus 40, see www.rcsb.org/pdb/molecules/pdb47_2.html.

Catabolite Activator Protein: a Second Messenger

**DECEMBER, 2003**—Bacteria love sugar. In particular, bacteria love glucose, which is easily digestible and quickly converted to chemical energy. When glucose is plentiful, bacteria ignore other nutrients in their environment, feasting on their favored source. But, when glucose is rare, they shift gears and mobilize the machinery needed to use other sources of energy.

Bacteria use an unusual modification of ATP, the molecule that carries chemical energy in the cell, to notify its synthetic machinery about what it is currently eating. As glucose levels drop, the cell-surface enzyme adenyl cyclase is activated. It grabs ATP molecules, clips off two phosphates, and reconnects the free end back onto the molecule, creating an odd little molecular loop through the phosphate. This product, called cyclic AMP, is released and it spreads through the cell, stimulating production of the enzymes that process other food molecules. Because of its role in delivering messages from the primary glucose sensor (adenyl cyclase) to the synthetic machinery, cyclic AMP is often known as a second messenger.

Catabolite activator protein (CAP), also known as cyclic AMP receptor protein (CRP), is activated by cyclic AMP and stimulates synthesis of the enzymes that break down non-glucose food molecules. It is composed of two identical subunits, shown in PDB entry **1cgp**. When cyclic AMP binds, it changes the conformation of the protein slightly, making it perfect for binding to DNA.

CAP binds to a specific DNA sequence, which is found next to the genes that are activated. When CAP binds to DNA, it coaxes RNA polymerase into place, beginning transcription.

For more information about the catabolite activator protein, see www.rcsb.org/pdb/molecules/pdb48_2.html. ◆
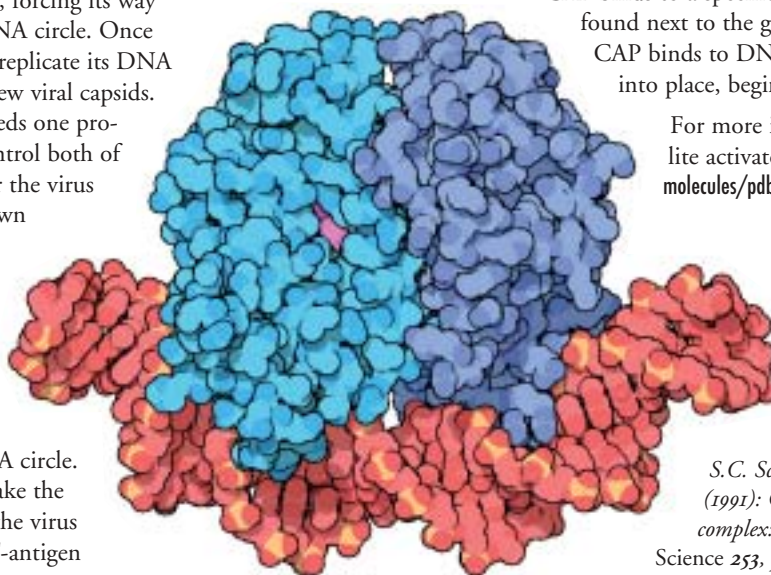
Edward (Ted) N. Baker is a Professor of Structural Biology at the University of Auckland in New Zealand. He is a member of the PDB Advisory Committee, and a long-time depositor to the PDB. Following a post-doctoral fellowship with Prof. Dorothy Hodgkin, in Oxford, he joined the staff at Massey University where he initiated a protein crystallography research program by determining the structure of the kiwifruit enzyme actinidin—the first protein structure to be determined in the Southern Hemisphere, and one of the first protein structures anywhere to be refined at high resolution. He is also responsible for the first crystallographic characterization of the milk protein, lactoferrin. In 1993 he was recognized as an International Research Scholar of the Howard Hughes Medical Institute. In 1997, he was awarded the Royal Society of New Zealand's Hector Medal in recognition of his innovation and leadership in studying the relationships between protein structure and function. He has served the community as President (1996-1999) of the International Union of Crystallography (IUCr), and played a leading role in developing accepted guidelines for the deposition of macromolecular data. He was involved in the creation of Acta Crystallographica Section D *and now serves as joint Editor.*

The RCSB PDB interviewed Professor Baker regarding his perspective on developments in crystallography and in the PDB:

RCSB PDB: How did you become interested in crystallography and protein structure, and how have you seen this field evolve since you began?

Prof. Baker: I became interested in crystallography because I loved the idea that you could see molecules—it seemed such a clear and exciting goal. I entered protein crystallography because my wise Ph.D. supervisor steered me towards Oxford for a Postdoc. This was just after the lysozyme structure had been solved and David Phillips' group had moved to Oxford. Fred Richards was in the lab (building his Richards box—"Fred's folly"), Chris Anfinsen was visiting, Guy and Eleanor Dodson, Tom Blundell, and Vijayan were there, and the insulin structure came out. All very exciting. When I was thinking of going back to New Zealand, knowing that I wanted to do protein crystallography, and also that I would have virtually no resources, Dorothy gave me wonderful advice: "If you really want to do it, just get started and it will work out in the end." It did.

The technical changes—from a time when crystals had to be at least 0.5 mm in size, data collection took months, and we built wire models by hand—have revolutionized every aspect of the field: vastly improved crystallization methods, crystal freezing, fast data collection, synchrotrons, computer graphics (thanks to Alwyn Jones), automated methods (SOLVE!), and refinement. But I think what is most exciting is the way the knowledge and use of macromolecular structure has become central to biology. No longer is protein crystallography an esoteric, if awe-inspiring, pursuit that consumed lots of money and produced remarkable

The structure of actinidin—the first protein structure determined in the Southern Hemisphere.

PDB ID: **2act**

E.N. Baker, E.J. Dodson (1980): Crystallographic refinement of the structure of actinidin at 1.7 Ångstroms resolution by fast Fourier least-squares methods. Acta Crystallogr., Sect. A *36, p. 559.*

understanding for a few proteins. Now it is central to drug development and it can transform a field (witness the MHC structure or the ribosome).

RCSB PDB: You were a member of the inaugural Editorial Advisory Board for *Acta D* and are currently joint Editor. Please tell us about the formation of *Acta D*, and how it has been evolving over the years.

Prof. Baker: *Acta D* was begun in the early 1990's in recognition of the great expansion of macromolecular crystallography that was then beginning. At that time very few of the biological journals published structural papers, though that has changed radically. In the past ten years we have seen a remarkable growth of interest in crystallographic methods, and a great period of methods development. *Acta D* has reflected this. The next wave is the huge increase in the numbers of experimentally-determined structures. We are already seeing large numbers of crystallization papers coming forward, and we think that these, and the structures that follow, point logically towards the establishment of a new electronic journal.

RCSB PDB: From your perspective as a both a depositor and a member of the PDB Advisory Committee: are you pleased with the current state of the PDB and what suggestions would you make as we move forward?

Prof. Baker: I am very happy with the current state of the PDB. There were concerns a few years ago as to whether the PDB could cope with the explosive growth in new structures. But I think the current speed with which new depositions are processed

and released has allayed those fears. And I am also very pleased that the problem of having large numbers of structures "on hold" has largely gone away—in part due to simple changes in the deposition defaults. Annotations and quality checks can always be done better, and will depend on better capturing of additional data.

What I think is the biggest challenge is to be able to make the structural data more accessible and meaningful for users. As a crystallographer I know very well which parts of my own structures are well-defined and which are not. I can assess other crystal structures quite well, too. The challenge is to express the indications that are given by data quality, electron density, B factors, occupancies, and correlation coefficients, in forms that can be intuitively understood by non-crystallographic users. Similar challenges exist for NMR structures.

RCSB PDB: Along with the MSD-EBI and the PDBj, we have just announced the formation of the Worldwide PDB (wwPDB). How do you, as a member of the international community of PDB users, view this agreement?

Prof. Baker: I applaud this unreservedly. Traditionally, crystallographers always viewed the PDB as "their" database, and viewed it as a single international resource. After all, they provided the data (this now includes NMR spectroscopists, of course). This feeling became muddied in the 1990's, around the time of the transfer of the PDB from Brookhaven to the RCSB, and one could even hear references to "the European PDB" and "the U.S. PDB". How did a New Zealander, far from both, fit in? Thankfully, this is now a thing of the past, and I hope that the right framework can now be developed for the long-term management and maintenance of this single Worldwide PDB. ✦

PDB Education Corner by Katherine Kantardjieff

PDB's Education Corner features a different teacher each quarter, offering an account of how he or she uses the PDB to educate students. This quarter's column is by Prof. Katherine Kantardjieff, Professor of Chemistry and Biochemistry at California State University Fullerton.

The California State University (CSU) is the largest, most diverse, and one of the most affordable university systems in the country. For the majority of students seeking baccalaureate education in California, as well as those seeking professional training, the CSU is the gateway institution, significantly impacting education and the economy of our state. The CSU campuses are predominantly undergraduate institutions, where a majority of undergraduates in the sciences conduct laboratory-based research as part of their baccalaureate degree requirements. My campus, CSU Fullerton, became the 12th state college in California to be authorized by the Legislature in 1957.
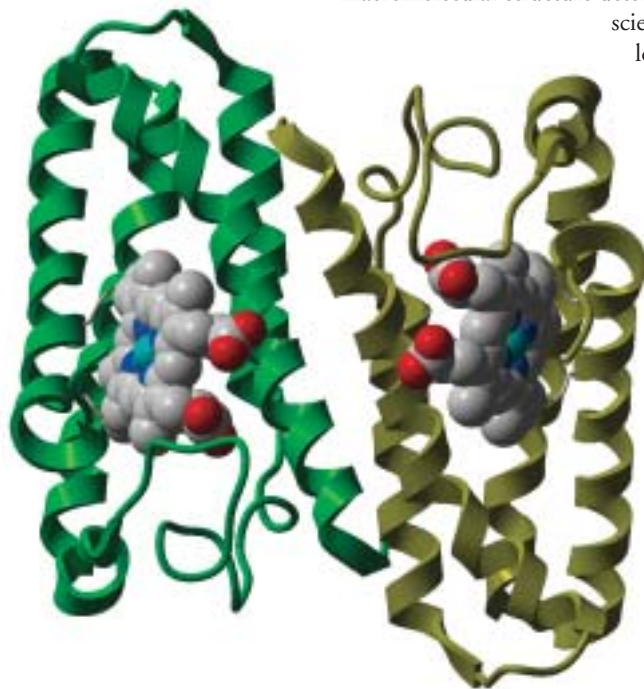
Today, CSUF has an enrollment of more than 32,000 students, making it the third largest in the 23-campus CSU system.

Protein Crystallography

Within the CSU, there is a vibrant California Program for Education and Research in Biotechnology (CSUPERB), which promotes system-wide biotechnology education and training, and supports several core research facilities. One such facility, based at CSUF, is the W.M. Keck Foundation Center for Molecular Structure (CMolS), which I direct. CMolS is the first comprehensive facility dedicated to research and education in both small and macromolecular structure determination and analysis using the science of crystallography, which is located at a predominantly undergraduate institution.

Undergraduates and Masters' level students learn about macromolecular structure determination methods in our classrooms and our training facilities, they learn about archiving and mining structural information in the Protein Data Bank, and they actually solve protein structures in our own research laboratories. Since 1997, our facilities have been available by remote access system-wide, and we annually host workshops for undergraduate faculty who wish to incorporate structure determination methods and molecular modeling into their curriculum.

Cytochrome c′ from Rhodobacter sphaeroides*, the first protein structure to be published in the* Journal of Chemical Crystallography. *This structure was the Masters' thesis of Prof. Kantardjieff's student, Laura Ramirez, who was supported by the NIH-MSD program. Ms. Ramirez is now Chair of the Science Department at East Los Angeles College. This image was rendered with ICM-Pro v3.0.24a, Molsoft LLC.*

PDB ID: **1gqa**

L.M. Ramirez, H.L. Axelrod, S.R. Herron, B. Rupp, J.P. Allen, K.A. Kantardjieff (2003): High resolution crystal structure of ferricytochrome c′ from Rhodobacter sphaeroides. Journal of Chemical Crystallography 33(5-6), pp. 413-24.

A Comprehensive Biochemistry Laboratory

A contemporary experiment in macromolecular structure/function analysis must cover not only advanced crystallographic techniques and methods, but also the front end aspects of protein crystallography, protein production, purification, and crystallization, as well as the back end aspects of structure validation and analysis. At CSU Fullerton, the majority of the upper division biochemistry laboratory (CHEM 422) is devoted to the study of the enzyme lactate dehydrogenase (LDH), from chicken breast muscle. With funding from CSUPERB, we have recently expanded the laboratory into a comprehensive structure determination and analysis center by adding an X-ray crystallography component.

LDH is a well-studied essential enzyme in carbohydrate metabolism for which extensive amino acid data are available for orthologous homologs, and for which some atomic resolution structure information is available (bacterial and human). In the laboratory, students isolate LDH-A from chicken breast muscle using standard techniques of tissue homogenization, centrifugation and ammonium sulfate precipitation, and LDH activity is later assayed spectrophotometrically using established protocols. Using their purified LDH-A, students gain experience in methods used in modern biotechnology to crystallize proteins for structure determination by X-ray diffraction analysis. This involves setting up 24-48 crystallization trials using commercially available random screens and vapor diffusion methods. In parallel, students also set up lysozyme crystallization trials using published procedures, and they are given an introduction to microbatch techniques, using saturated sodium chloride solutions under paraffin or mineral oil. Our students have succeeded in producing crystals of LDH-A from chicken breast muscle under a variety of screening conditions not previously reported for crystallizing either bacterial or human LDHs. We hope to have diffraction data collected before the end of this fall term, with the ultimate goal of making these data available for any biochemistry laboratory course. Students would be able to solve the structure by molecular replacement, and build models into electron density using available software, such as XtalView.

To complement the wet lab work, we have the students spend several laboratory sessions conducting related bioinformatics exercises that include visualization of the atomic resolution structural details of LDH homologs in the PDB. Students examine the details of the active site chemistry, looking at substrate and cofactor binding interactions and, using either DeepView or ICM-Pro, students make a homology model of their enzyme from chicken and conduct *in silico* mutagenesis experiments. Based on students' laboratory reports and general feedback, these bioinformatic and computational exercises using information in the PDB have greatly enhanced our students' understanding of protein biochemistry.

Contemporary Biology and the Art of Science

The Department of Biological Science at CSUF has designed and implemented a new curriculum that builds on a core program with themes and perspectives to connect and integrate major concepts, principles and facts. In one of the four freshman core courses, Cellular Basis of Life (BIOL 172), Dr. MerriLynn Casem has integrated the PDB into her teaching through use of the *Art of Science* exhibit, which CSUF has hosted during the Fall 2003

semester. In the beginning, Dr. Casem asked her students to view the images strictly as art, in whatever initial context the students brought with them. She asked them to critique the works first from an artistic perspective, and then with regard to information conveyed, to determine whether the models and graphics informed the students' understanding. Dr. Casem has since used the exhibit and the PDB as "an excellent reference point" for the themes emphasized in the course: order and organization, properties of life, and energy usage. The students have been particularly impressed by the fact that cells are not at all "empty", but rich with the molecules of life.

What we have been surprised and pleased to note is that, in addition to our own biochemistry and biology majors, students from outside the College of Natural Science and Mathematics have been sent to view the exhibit as part of their coursework, including student teachers and art majors. Thus, at CSUF, the PDB is being used to educate the broader campus community about molecular science.

Structural Bioinformatics

As part of the CSU mission to strengthen the California workforce, CSUF Extended Education offers many courses for continuing development of working professionals, including several certificate programs. Our Certificate in Bioinformatics, one of two granted system-wide, is distinguished by its capstone advanced course in sequence, structure and function analysis, which makes extensive use of structural information in the PDB. Students learn about structure-guided drug-design and explore protein-protein interactions. The Pasadena Bioscience Center—a joint endeavor between California State University, Caltech, Huntington Medical Research Institute, Pasadena City College, City of Pasadena, and bioscience industry representatives—also offers continuing education courses in protein structure and drug design, which make extensive use of the PDB. ✦

Related Links: FTP Resources

The RCSB PDB offers resources for using the PDB FTP site. These include several scripts for mirroring and automated downloading of files:

getPdbUpdate
ftp://ftp.rcsb.org/pub/pdb/software/getPdbUpdate.html
Perl script to retrieve files from any update found at
ftp://ftp.rcsb.org/pub/pdb/data/status/

getPdbStructures
ftp://ftp.rcsb.org/pub/pdb/software/getPdbStructures.html
Perl script to upload a list of PDB ID's and retrieve data files for those entries

rsyncPDB
ftp://ftp.rcsb.org/pub/pdb/software/rsyncPDB.sh
Script to set up a local PDB FTP server; general rsync documentation can be found at www.samba.org/rsync

RCSB PDB Job Listings

RCSB PDB career opportunities are posted at www.rcsb.org/pdb/jobs.html.
The current available openings are:

Structural Bioinformatics Project Leader

The Protein Data Bank (PDB) group at the University of California, San Diego is seeking a Project Manager to guide the PDB in its next five-year phase of development. The Project Manager will work collaboratively with the PDB software architects, programmers, and scientists, at UCSD and the RCSB PDB partner sites, to expand the PDB's functionality and reliability as a premier biological data and information resource.

Job functions include: identify and develop requirements for new PDB delivery, query functionality and usability; develop and implement innovative approaches that will satisfy users' current needs and anticipate their future needs based on progress in the science of structural biology and structural bioinformatics; and work with the scientific community to fulfill the above.

Qualifications:

- Ph.D. in biological sciences or related field or equivalent combination of education and experience in the field of bioinformatics and computational biology, including expert knowledge and research experience in sequence and protein structure analysis, protein 3-D structure prediction and fold recognition, and protein modeling.
- Extensive background and expertise in project management and research coordination.
- Demonstrated experience working with a team of high-level professional scientists and computer programmers.

- Advanced skills and experience in developing biological databases and in SQL.
- Ability to communicate and deal effectively and productively with people at all levels of responsibility in various functional areas.
- Demonstrated ability to work with a diverse set of people to solve problems and build consensus.
- Strong, demonstrated experience in software development, especially with Enterprise Java and multi-tier architectures.

Applicants should apply on-line at joblink.ucsd.edu/bulletin/job.html?cat=new&job_id=31324.

Biochemical Information Specialist

The Protein Data Bank at Rutgers University has a position open for a Biochemical Information Specialist to curate and standardize macromolecular structures for the Protein Data Bank. A background in biological chemistry, as well as some experience with UNIX-based computer systems, is required. Experience in crystallography and/or NMR spectroscopy is a strong advantage. The successful candidate should be well-motivated, able to pay close attention to detail, and meet deadlines. This position offers the opportunity to participate in an exciting project with significant impact on the scientific community.

Please send resume to Dr. Helen Berman at pdbjobs@rcsb.rutgers.edu.

RCSB PDB Leadership Team

The overall operation of the PDB is managed by the RCSB PDB Leadership Team.
Technical and scientific support are provided by the RCSB PDB Members.

Dr. Helen M. Berman
Director
Rutgers University
berman@rcsb.rutgers.edu

Dr. Philip E. Bourne
Co-Director
SDSC/UCSD
bourne@sdsc.edu

Judith Flippen-Anderson
Production and Outreach Leader
Rutgers University
flippen@rcsb.rutgers.edu

Dr. Gary Gilliland
Co-Director
CARB/NIST
gary.gilliland@nist.gov

Dr. John Westbrook
Co-Director
Rutgers University
jwest@rcsb.rutgers.edu

RCSB PDB Members

The RCSB PDB is managed by three 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

PRENTICE BISBAL
prentice@rcsb.rutgers.edu

KYLE BURKHARDT
kburkhar@rcsb.rutgers.edu

LI CHEN
lchen@rcsb.rutgers.edu

SHARON COUSIN
sharon@rcsb.rutgers.edu

DR. SHUCHISMITA DUTTA
sdutta@rcsb.rutgers.edu

DR. ZUKANG FENG
zfeng@rcsb.rutgers.edu

LEW-CHRISTIANE
FERNANDEZ
fernandz@rcsb.rutgers.edu

DR. RACHEL KRAMER
GREEN
kramer@rcsb.rutgers.edu

VLADIMIR GURANOVIC
vladimir@rcsb.rutgers.edu

DR. SHRI JAIN
sjain@rcsb.rutgers.edu

DR. ROSE OUGHTRED
rose@rcsb.rutgers.edu

DR. IRINA PERSIKOVA
persikova@rcsb.rutgers.edu

SUZANNE RICHMAN
richman@rcsb.rutgers.edu

MELCOIR ROSAS
melcoir@rcsb.rutgers.edu

DR. BOHDAN SCHNEIDER
bohdan@rcsb.rutgers.edu

DR. HUANWANG YANG
hyang@rcsb.rutgers.edu

DR. JASMIN YANG
jasmin@rcsb.rutgers.edu

CHRISTINE ZARDECKI
zardecki@rcsb.rutgers.edu

SDSC/UCSD

San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0537

DR. KEN ADDESS
addess@sdsc.edu

DAVID ARCHBELL
dave@sdsc.edu

TAMMY BATTISTUZ
tammyb@sdsc.edu

DR. WOLFGANG F. BLUHM
wbluhm@sdsc.edu

DR. NITA DESHPANDE
nita@sdsc.edu

JEFF MERINO-OTT
jott@sdsc.edu

WAYNE TOWNSEND-MERINO
wayne@sdsc.edu

CARB/NIST

Center for Advanced Research in
Biotechnology
National Institute of Standards and
Technology
9600 Gudelsky Drive
Rockville, MD 20850

AL CARLSON
carlson@umbi.umd.edu

DR. VEERASAMY RAVICHANDRAN
vravi@nist.gov

KATHRYN ROSECRANS
rosecran@umbi.umd.edu

ELIZABETH WALKER
walkere@umbi.umd.edu