New Routes to Crystallographic Data Publication:

The Protein Data Bank in an Open Data Community
&
The Protein Structure Initiative Structural Genomics Knowledgebase

John Westbrook

IUCr 2008 OSAKA

Introduction

- Content of the archive
- Highlights of the 37 year history
- PDB worldwide management organization
- Data standards
- Data deposition and archive management
- Enabling other integrated forms of delivery
- New project - PSI SGKB

What is the PDB?

- Single international repository for all information about the structure of large biological molecules
- Archival database with hundreds of thousands of users who depend on the data

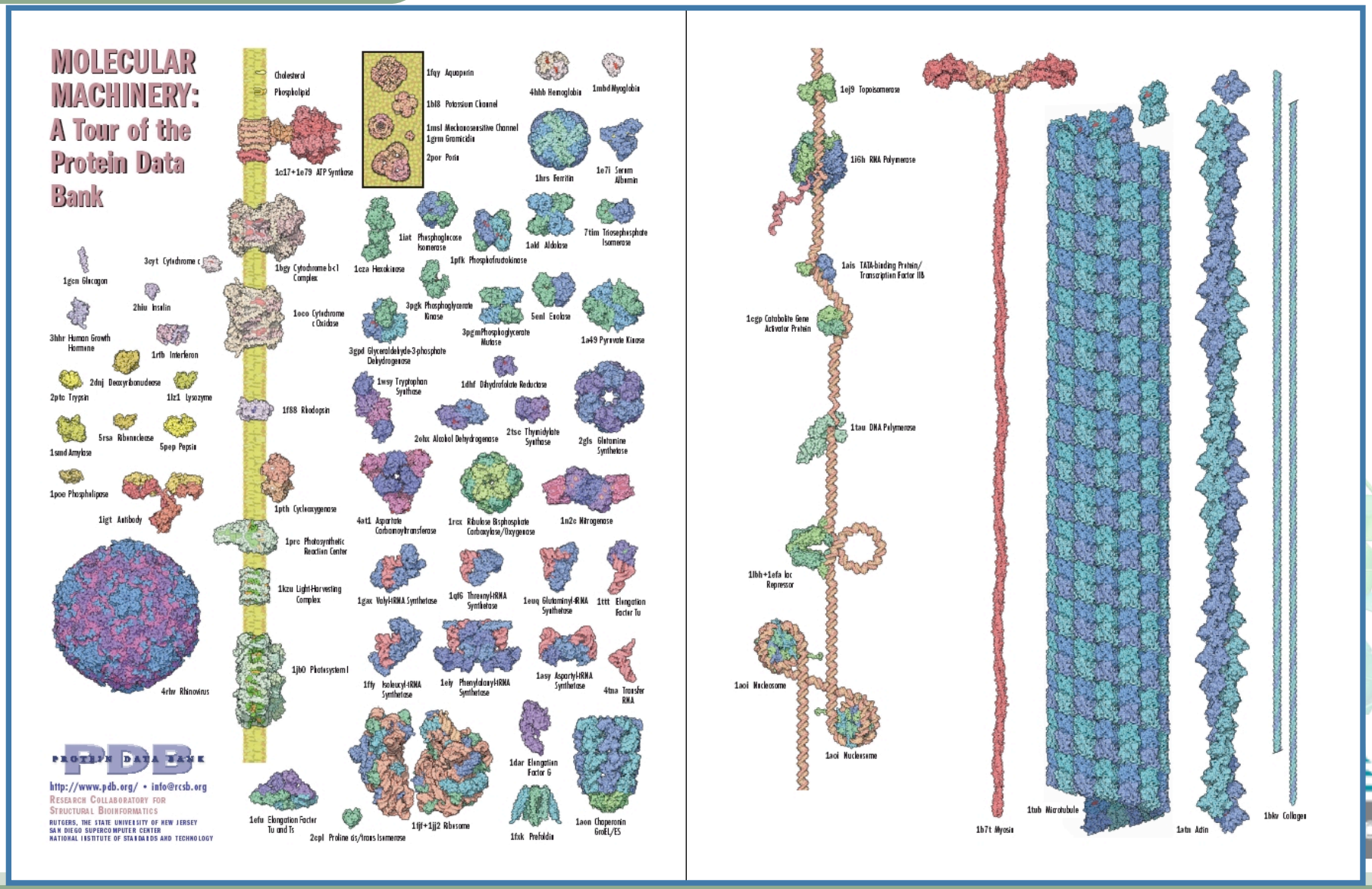
Archive Contents

- Public archive
  - More than 300,000 files in PDB, mmCIF, & PDBML/ XML formats (Jan 2008)
  - Requires over 70 GBbytes of storage
  - Data & chemical dictionaries
  - Derived data files
- For each entry
  - Atomic coordinates
  - Sequence information
  - Description of structure
  - Experimental data
  - Release status information
- Internal archive
  - Depositor correspondence
  - Depositor contact information
  - Paper records
  - Documentation
  - Historical records from Day One

WORLDWIDE PDB PROTEIN DATA BANK

Worldwide Protein Data Bank

www.wwpdb.org



History of the Archive

1970s

- Community discusses how to create archive protein structures
- Cold Spring Harbor meeting in protein crystallography
- PDB established at Brookhaven (Oct 1971; 7 structures)

1980s

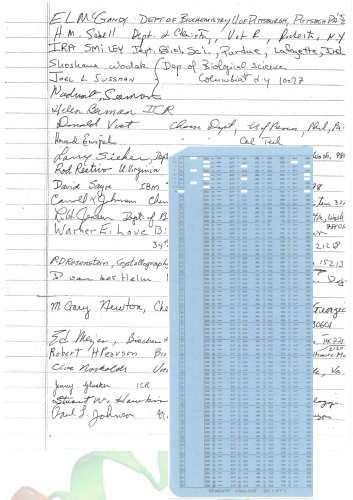
- Number of structures increases as technology improves
- Community discussions about requiring depositions
- IUCr guidelines established
- Number of structures deposited increases

1990s

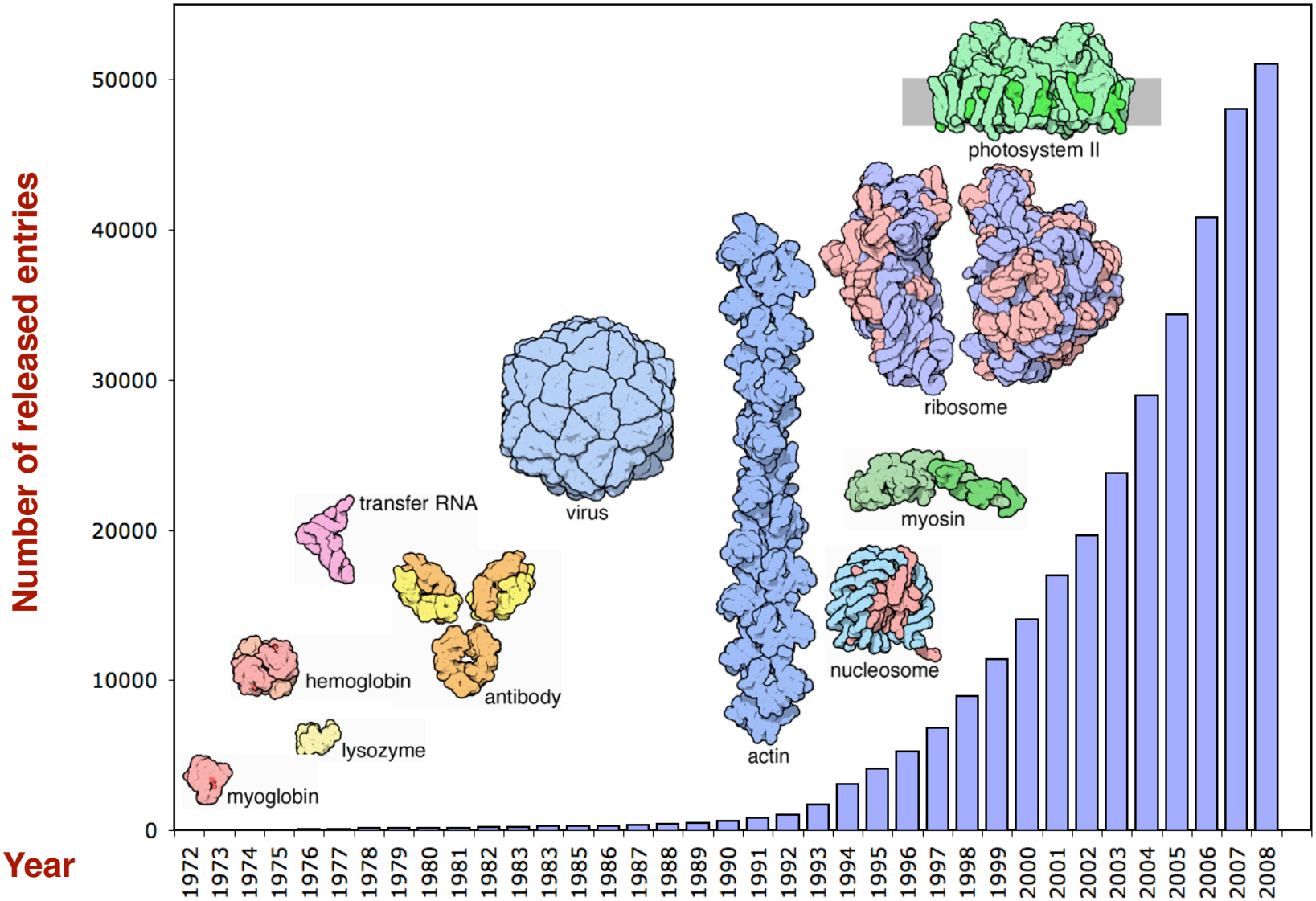
- Structural genomics begins
- PDB moves to RCSB PDB

2000s

- wwPDB formed
- Mandatory deposition of structure factors and restraints (Feb 2008)
- > 24 Journals require PDB IDs for publication (2008)
- 50,000th structure released (April 2008)



IUCr 2008 OSAKA


Number of released entries
50000
40000
30000
20000
10000
0
Year
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1983
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
photosystem II
ribosome
virus
transfer RNA
myosin
nucleosome
hemoglobin
antibody
lysozyme
actin
myoglobin

wwPDB

- Members
  - RCSB PDB (Research Collaboratory for Structural Bioinformatics)
  - PDBj (Osaka University)
  - PDBe (EMBL-EBI)
  - BioMagResBank
- MOU signed July 1, 2003
- Announced in *Nature Structural Biology* November 21, 2003

wwPDB Resources

WORLDWIDE PDB PROTEIN DATA BANK

Welcome to the Worldwide Protein Data Bank

Home | wwPDB Agreement | Statistics | FAQ | News | Contact Us

- Access the PDB FTP:
  - RCSB PDB | PDBe | PDBj
  - Archive Download
- Deposit Data to the PDB:
  - RCSB PDB | PDBe
  - PDBj | BMRB
- Search wwPDB Websites:
  - RCSB PDB | PDBe
  - PDBj | BMRB
- PDB Archive Snapshots
- Instructions to Journals
- PDB Remediation
  - Description
  - Chemical Component Dictionary
  - Software
- Documentation
  - Format
  - Annotation
  - Remediation
- Workshops
  - X-ray Validation

The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for PDB data. The founding members are **RCSB PDB** (USA), **PDBe** (Europe) and **PDBj** (Japan)'. The **BMRB** (USA) group joined the wwPDB in 2006. The mission of the wwPDB is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community.

This site provides information about services provided by the individual member organizations and about projects undertaken by the wwPDB.

Please note: ftp://ftp.rcsb.org is no longer updated. Please access the PDB archive using one of the FTP sites listed in the left menu.

14-August-2008

IUCr: wwPDB Exhibition Stand and Pres

The wwPDB partners will be exhibiting at the XX
International Union of Crystallography (IUCr; Aug
#14. Please stop by for website demonstrations
around the globe.

Helen M. Berman (RCSB PDB) will present a key
Data Bank tells us about the past, present, and f
August 24.

On Saturday, August 30, John Westbrook (RCSB
PDB Archive".

11-August-2008

Download Statistics Available by Structu

Downloads from the PDB archive are one of the primary means of accessing scientific structure results. While there are cross-links between the corresponding scientific publication and the PDB entry, in many cases it is the structure file that is accessed and downloaded more frequently.

www.wwpdb.org

Guidelines and Responsibilities

- All members issue PDB IDs and serve as distribution sites for archival data files
- One member is the archive keeper (RCSB PDB)
- All format and data dictionary documentation publicly available
- Strict rules for redistribution of PDB data files
- All sites can create their own delivery websites

Common Data Standards

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

PDB Home | Contact Us

Dictionary Home | PDBML Home | Software Tools Home

PDBML Resources

PDBML

PDBML

The Protein Data Bank Markup
provided in XML schema of the
Exchange Data Dictionary Othe
also presented in the list below

- PDBML data files are pro
  - fully marked-up file
  - files without atom
  - files with a more s
- Data files in PDBML form
- Software tools for manip
- An article describing PDE
  PDBML: the representati
  John Wesbrook, Nobutos
  Bioinformatics, 21(7), 98

PDBML Schema

- **PDB Exchange Diction**
  XML Schema for Exchan
  between MSD-EBI, PDBj

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

PDB Home | Contact Us

Dictionary Home | PDBML Home | Software Tools Home Search: PDB Exchange Dictionary Go

Dictionary Resources

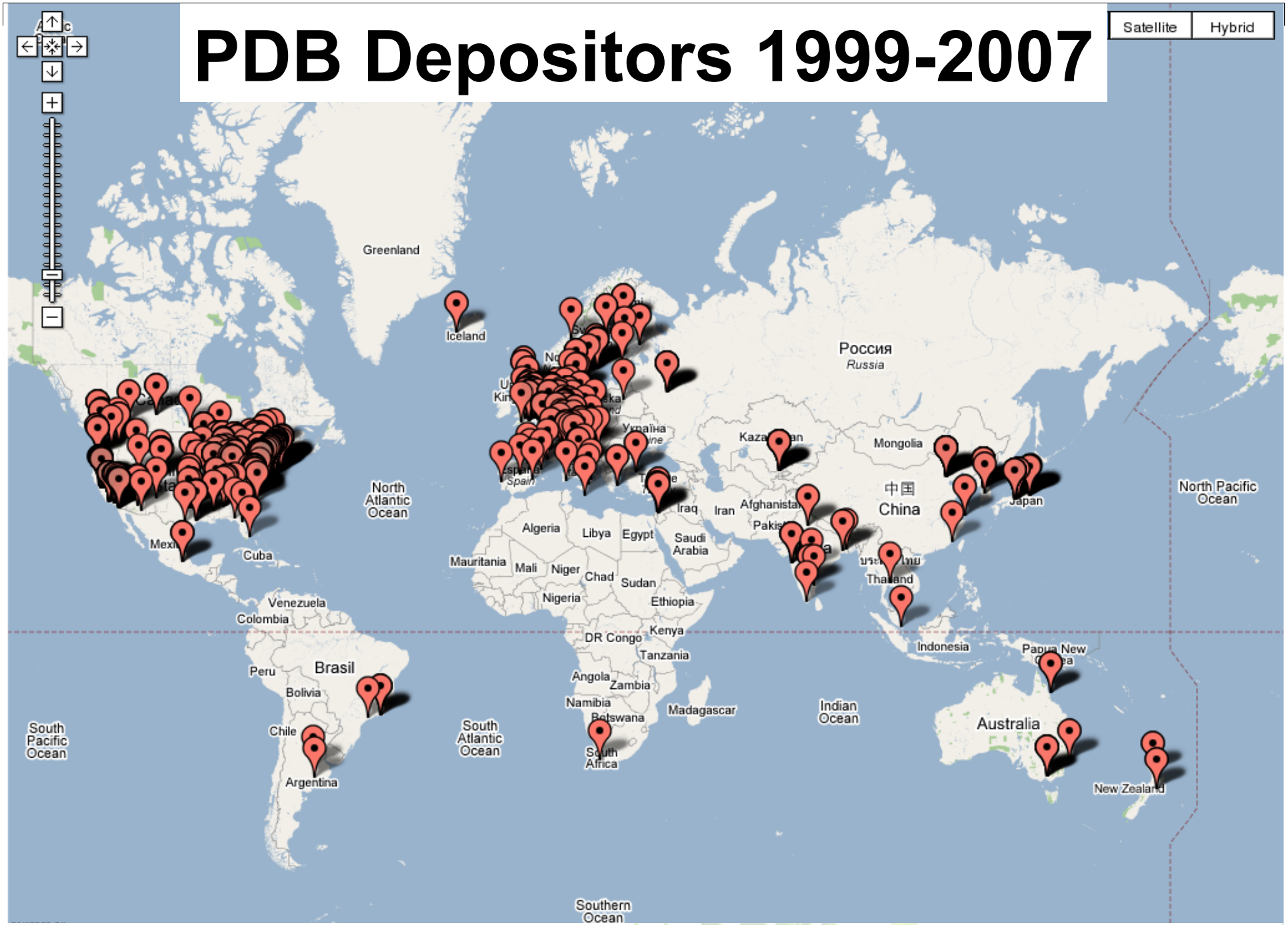
The Protein Data Bank (PDB) uses macromolecular Crystallographic Information File (mmCIF) data dictionaries to describe the information content of PDB entries. The PDB Exchange data dictionary consolidates content from a variety of crystallographic dictionaries including: the IUCr Core, mmCIF, Image and symmetry dictionaries. The PDB Exchange Dictionary also includes extensions describing NMR, Cryo-EM, and protein production data. PDB data processing, data exchange, annotation, and database management operations all make heavy use of the data format and the content of the PDB Exchange Dictionary. Software tools are used to convert mmCIF data files to the older PDB format and to PDBML/XML.

- Data files in mmCIF format can be downloaded from the RCSB PDB website or by ftp.
- Software tools are available for preparing and editing depositions.
- Software tools are available for converting mmCIF data files to PDB and PDBML formats
- A complete list of PDB software tools for managing PDB data in mmCIF format can be found here.

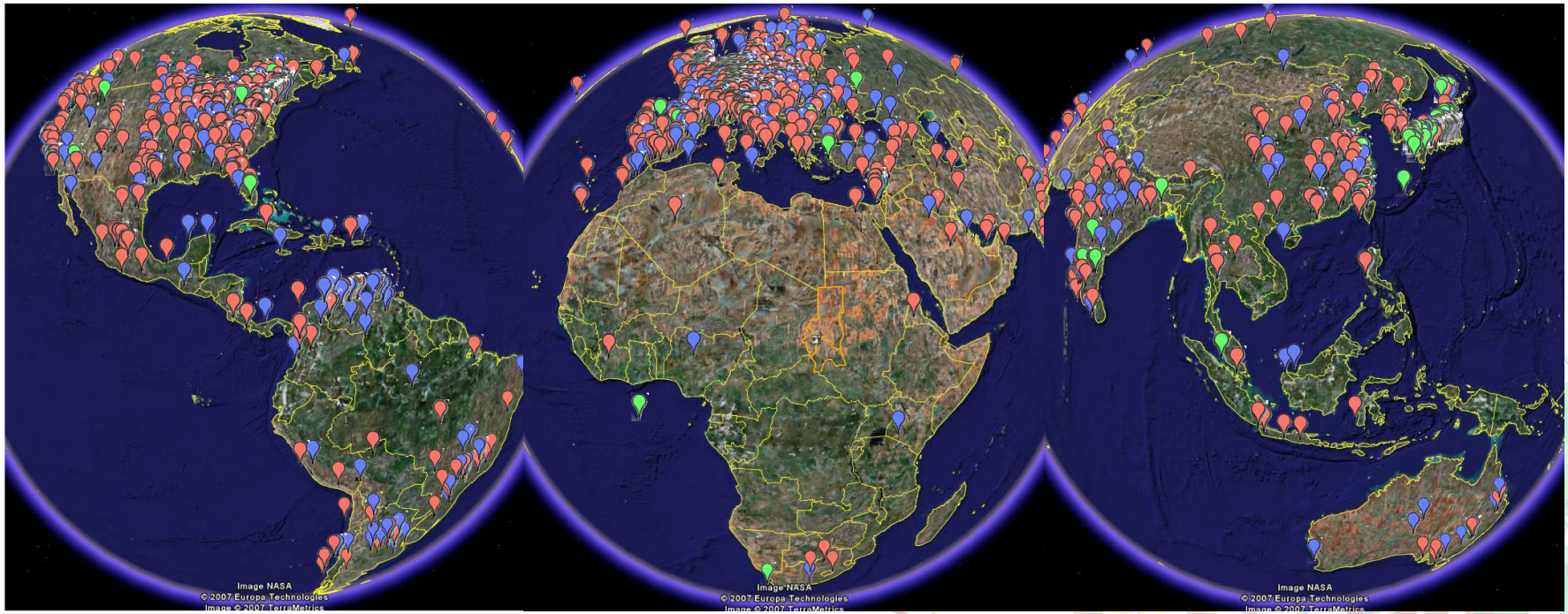
Dictionary Content and Representation

- Background and Introduction about mmCIF
- Chapter 3.6. Classification and use of macromolecular data. (PDF) in *International Tables for Crystallography G.* Definition and exchange of crystallographic data, S.R. Hall and B. McMahon, Editors. 2005, Springer: Dordrecht, The Netherlands. p. 144-198.
  - Appendix 3.6.2 The Protein Data Bank exchange dictionary (PDF) in *International Tables for Crystallography G.* Definition and

PDB Depositors 1999-2007



PDB FTP Traffic



RCSB PDB
164.5 million
data downloads

MSD-EBI
23.8 million
data downloads

PDBj
9.5 million
data downloads

Statistics on PDB FTP Users

- Approximately 200,000,000 data files were downloaded in 2007
- This does not include redistribution (by other resources, such as PDBsum, OCA, Jena, Relibase, Accelrys)
- Also does not include extensive, albeit anecdotal, use of PDB behind the firewalls of pharmaceutical and biotechnology companies

Time-stamped copies of the archive

- 57 Gbytes of data for 2006, released January 2, 2007
- 68 Gbytes of data for July 2007 snapshot
- Both include
  - PDB format entries
  - mmCIF format entries
  - PDBML format entries
  - Experimental data
  - Dictionary, schema, and format documentation

PDB IDs and DOIs

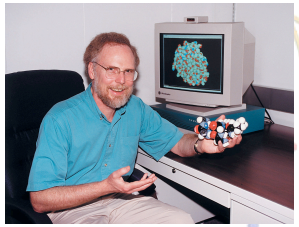
- Digital Object Identifiers (DOIs) created for each PDB entry
- DOIs provide credit for a PDB entry in CVs
- Used as a reference in publications -

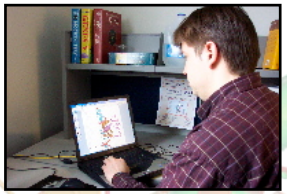
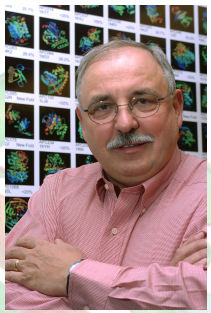
  http://dx.doi.org/10.2210/pdb4hhb/pdb

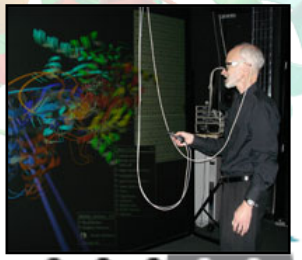
Community

- **Depositors**
  - Different methods: X-ray, NMR, cryo-EM
- **Users**
  - Structural biologists
  - Biochemists
  - Molecular biologists
  - Computational biologists
  - Educators
  - Students
  - Lay community



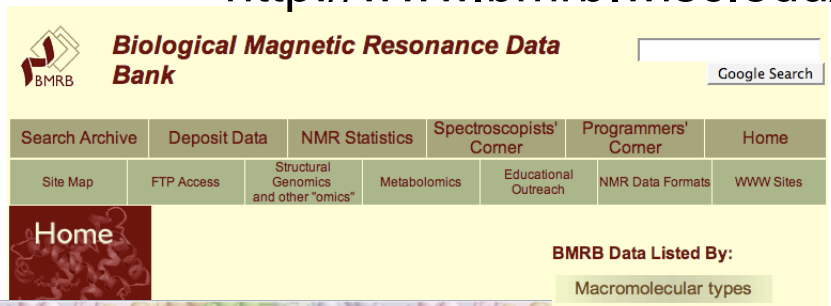



Worldwide Protein Data Bank

www.wwpdb.org

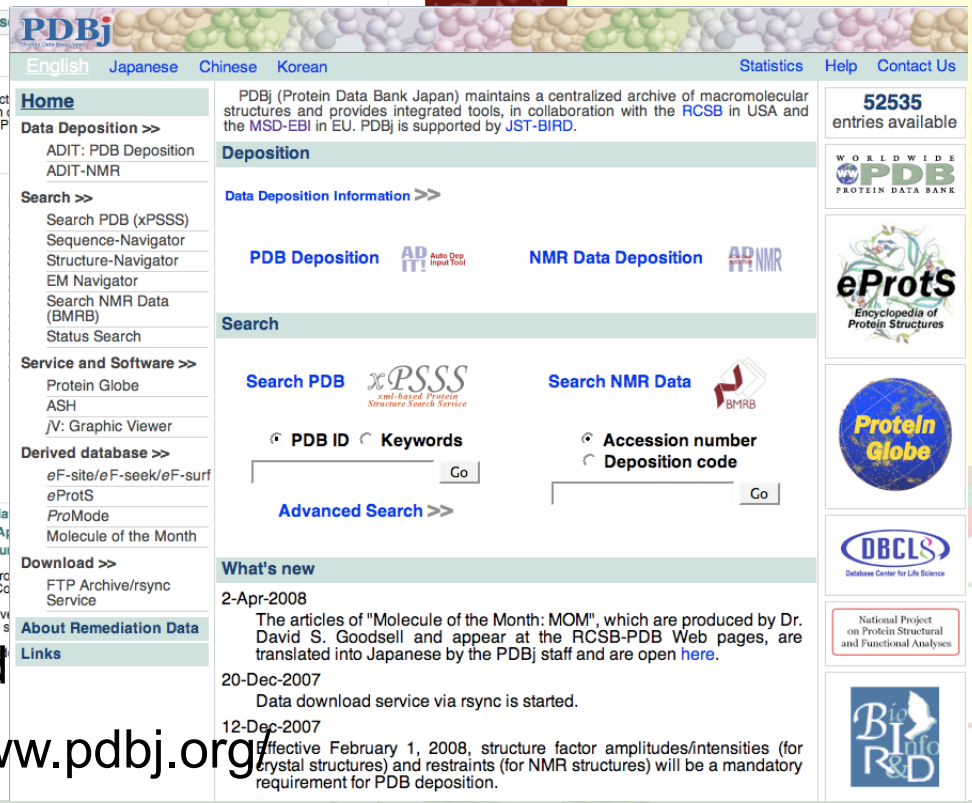
http://www.rcsb.org/pdb/

http://www.bmrb.wisc.edu/



http://www.ebi.ac.uk/msd

http://www.pdbj.org/



IUCr 2008 OSAKA

wwPDB Funding

RCSB PDB PROTEIN DATA BANK

NSF, NIGMS, DOE, NLM, NCI,
NCRR, NIBIB, NINDS, NIDDK

RUTGERS UCSD

PDBe PROTEIN DATA BANK EUROPE

Wellcome Trust, EU,
CCP4, BBSRC, MRC, EMBL

PDBj Protein Data Bank Japan

BIRD-JST, MEXT

BMRB

NLM

IUCr 2008 OSAKA

Knowledgebase Vision

The PSI Structural Genomics Knowledgebase (PSI SG KB) will turn the products of the PSI effort into major advances in knowledge that can be used to understand living systems and human disease.

The PSI SG KB will be a key resource for the advancement of biology, biochemistry, functional genomics, pharmacology, bioinformatics, chemistry, education and clinical medicine.

Poster – P04.22.430*(C364)*

PSI SGKB

PSI SGKB - http://kb.psi-structuralgenomics.org

PSI Protein Structure Initiative

PSI Structural Genomics Knowledgebase

HOME | ABOUT PSI | ABOUT PSI SGKB | OUTREACH/EDUCATION | BIOMEDICAL THEMES | CONTACT US

PSI-2 CENTERS
- Large Scale Centers
- Specialized Centers
- Modeling Centers
- Resource Centers

KB MODULES
- Target Selection
- Experimental Data Tracking
- Materials Repository
- Model
- Annotation
- Metrics
- Technology

GETTING STARTED

WHAT'S NEW?

SITEMAP

ACCESSIBILITY

Welcome to the PSI SGKB Portal!

To get started please read our FAQ. Please send comments to comments@psi-structuralgenomics.org

The Protein Structure Initiative Structural Genomics Knowledgebase (PSI SGKB) is designed to turn the products of the Protein Structure Initiative effort into knowledge that is important for understanding living systems and disease. The PSI SGKB is a key resource in the advancement of biology, biochemistry, functional genomics, pharmacology, bioinformatics, education and clinical medicine.

Use the following query form to search the PSI SGKB by one-letter code protein sequence, keyword or Protein Data Bank (PDB) identifier code. (Example Sequence Search)

Sequence Search
Keyword Search
PDBId Search

Go | Reset | Help

NEWS

Web Portal to Advance Wide Range of n Studies
otein Structure Initiative (PSI), an effort ted by the National Institutes of Health (NIH), has launched an online resource that will enable scientists from across biomedical disciplines to easily access a wealth of information about proteins and to sp ... Read more »

FEATURED PSI STRUCTURES

SARS Coronavirus Nonstructural Protein 1
Researchers at the Joint Center for Structural Genomics have obtained the first look at nsp1 (nonstructural protein 1), a major factor in the pathogenicity of the coronavirus that causes SARS (severe acute respiratory syndrome).... Read more »

FUNCTIONAL SLEUTH

ional Sleuth presents PSI structures lacking full ional annotation. Explore these protein structures dd your input about the possible functions Begin ring »

TECHNICAL HIGHLIGHT

Model Validation & Fold Function Analysis
MCSG tries to automate the process of fold recognition, assignment and model validation, automate deposition of structures to databases, and speed up data release. Development a database to monitor and evaluate all steps of the process from gene to structure is essential. The database creates a self-training system to aid decision making process and improve ... Read more »

Search by
- Sequence
- Keyword
- PDB ID

Newly released PSI-solved structures

News and Events

Molecules of Unknown Function

Link to the Functional Sleuth Gallery

Featured Structure

Technology Features

Link to the Technology Module

Integrates query results across the pipeline

PSI Protein Structure Initiative

PSI Structural Genomics Knowledgebase

HOME | ABOUT PSI | ABOUT PSI SGKB | OUTREACH/EDUCATION | BIOMEDICAL THEMES | CONTACT US

PSI-2 CENTERS
- Large Scale Centers
- Specialized Centers
- Modeling Centers
- Resource Centers

KB MODULES
- Target Selection
- Experimental Data Tracking
- Materials Repository
- Model
- Annotation
- Metrics
- Technology

GETTING STARTED

WHAT'S NEW?

SITEMAP

ACCESSIBILITY

SUMMARY | DB REPORT

Structures | Annotations | Models | Targets | Protocols | Materials

TOTAL 18 TARGET(S) FOUND GLOSSARY OF TERMS

TargetDB : GO.6759 | View alignment (I = 100%
| Target Status : Other | Source organism : Arabi
Category : legacy - no

TargetDB : GO.6760 | View alignment (I = 100%
| Target Status : Other | Source organism : Arabi
Category : legacy - no

TargetDB : GO.74365 | View alignment (I = 100
| Target Status : Other | PDB ID : 2I9Y | Source
Target Category : legacy - no

TargetDB : GO.33961 | View alignment (I = 100
| Target Status : Other | Source organism : Arab
Category : legacy - Unknown selection

TargetDB : GO.6764 | View alignment (I = 92%)
| Target Status : Other | Source organism : Arabi

PSI Protein Structure Initiative

An Information Portal to Biological Macromolecular Structures

Print Page | New Search

PSI Knowledgebase | TargetDB Home | PepcDB Home | Protein Data Bank

TargetDB | Target Query Results

There is 1 sequence that match your request.

ID: GO.74365	*Lab:* CESG *Latest update:* 2005-04-29
Name	At1g70830.1
Status	Selected,Cloned,Expressed,Soluble,NMR Assigned,HSQC,NMR Structure,In PDB,Other
Database Reference	PDB: 2I9Y
Source Organism	Arabidopsis thaliana
Sequence	TEASSLVGKLETDVEIKASADKFHHMFAGKPHHVSKASPGNIQGCDLHEGDWGTVGSIVFWNYVHDGEAKVAKERI EAVEPDKNLITFRVIEGDLMKEYKSFLLTIQVTPKPGGPGSIVHWHLEYEKISEEVAHPETLLQFCVEVSKEIDE HLLAEE
Domain Annotation	Pfam
Other Sources	Superfamily | TIGR Families | ProDom | iProClass | Prosite

Protein Properties

Number of Residues	Mol. Weight	Avg. Hydropathy Score	Charge	Ip Value
157	17529	-0.339	-8.0	4.9

PSI SGKB

New PSI-Nature SG KB Homepage

PSI | nature
StructuralGenomicsKnowledgebase

Home
Structural Genomics Update
About this site
About PSI
PSI Centers
PSI Resources
NPG Resources

Welcome to the
Structural Genomics Knowledgebase

The Protein Structure Initiative - Nature Structural Genomics Knowledgebase is designed to turn the products of the Protein Structure Initiative effort into knowledge that is important for understanding living systems and disease.

Use this site to explore the PSI's work and to stay informed about advances in structural genomics and structural biology.

search — Explore proteins and the Knowledgebase here

by sequence
by plain text
by structure (PDB id)
example query
help
search

Structural Genomics Update September 2008
Research advances, news and events in Structural Biology

Featured molecule

Aspartate Dehydrogenase

Aspartate dehydrogenase is a new enzyme discovered by the Joint Center for Structural Genomics and the Northeast Structural Genomics Consortium. The path to discovery began with the genome of the bacterium...

more

Research advances

featured article
Nunc pulvinar tincidunt
article ref and date

News

New Web Portal to Advance Wide Range of Protein Studies

e-alerts
Receive monthly updates by email.
sign up

RSS (monthly updates)

RSS (new molecules)

functional sleuth

Functional Sleuth presents PSI structures lacking full functional annotation.

begin exploring

latest PSI statistics

New structures last month: 259
Total structures to date: 3055
Total distinct structures: 1584

more

see latest structures

Coming Soon …

PSI SGKB

Acknowledgements

Director – Helen Berman

KB Team

Wendy Tao
Raship Shah
James Chun
Margaret Gabanyi

Modules

Torsten Schwede (Models)
Andrei Kouranov (Exp. Data Tracking)
Paul Adams (Technology)
Wladek Minor (Publications)
Josh La Baer (Materials)
Rajesh Nair (Metrics)

Access Information

http://kb.psi-structuralgenomics.org

Funding for the PSI SG KB

National Institute of General Medical Sciences

PSI SGKB