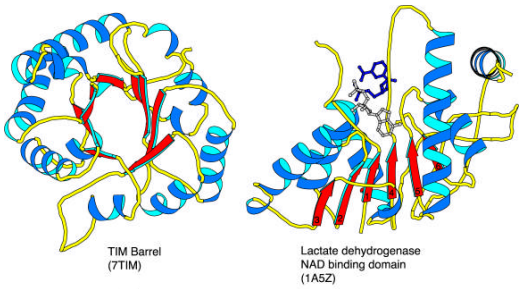
Protein Structure Determination



How are these structures determined?

Why Bother With Structure?

- The amino acid sequence of a protein contains interesting information.
- A protein sequence can be compared to other protein sequences to establish its evolutionary relationship to other proteins and protein families.
- However, for the purposes of understanding protein function, the 3D structure of the protein is far more useful than the sequence.

Protein Sequences Far Outnumber Structures

- Only a small number of protein structures have been experimentally determined.

  PDB ~49,760 protein structures
  Genebank ~61,132,599 sequences

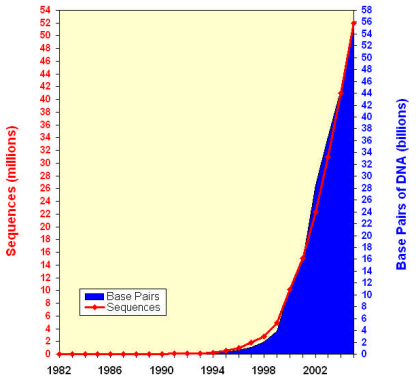
- Of the 49,760 structures, only **7997** are dissimilar in sequence (<30% ID).

Growth of GenBank

Now over 61M sequences and 65B base pairs from 200K organisms

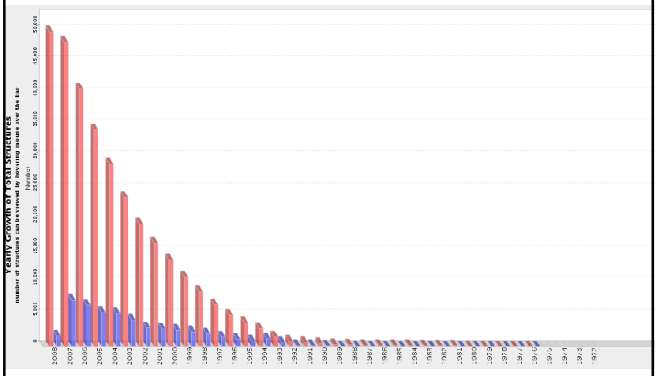
(1982 - 2005)

Exponential Increase



http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

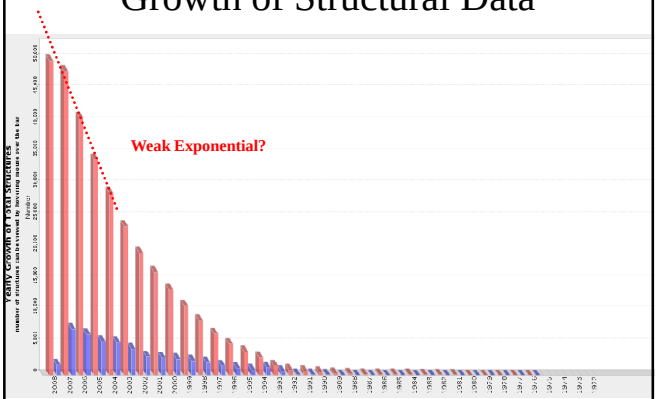
Growth of Structural Data



http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100

Growth of Structural Data

Weak Exponential?



Structural Proteomics

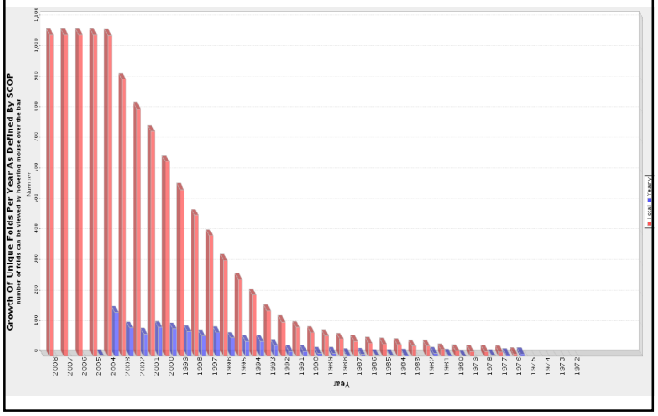
- Use experimentally determined structures to model the structures of similar proteins
  - Threading
  - Homology Modeling
  - Fold recognition

  } **Avoids *Ab initio* structure determination**
- Need representative protein structures for the total repertoire of protein folds
- Provide 3D portraits for all proteins in an organism
- Goal: Use structure to infer function.
  - More sensitive than primary sequence comparisons

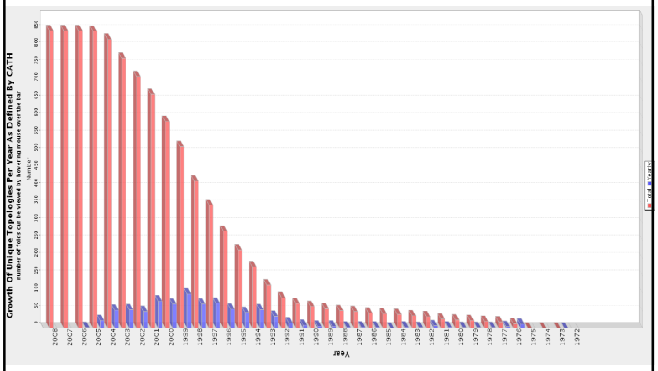
Redundancy in PDB (1 April 08)

Sequence identity	Number of non-redundant chains
90%	17384
70%	15427
50%	12818
30%	7997

Unique folds in PDB



Unique topologies in PDB



New Topologies and Folds Becoming Rare

Structural Genomics

PSI Protein Structure Initiative

Initiated in 1999 by NIH
Phase I included 9 large centers for high throughput structure determination
Phase I ran from ~2000 - 2005

Goal

The long-range goal of the Protein Structure Initiative (PSI) is to make the three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences.

http://www.nigms.nih.gov/psi/mission.html

Structural Genomics

Benefits

Structural descriptions will help researchers illuminate structure-function relationships and thus formulate better hypotheses and design better experiments.

The PSI collection of structures will serve as the starting point for structure-based drug development by permitting faster identification of lead compounds and their optimization.

The design of better therapeutics will result from comparisons of the structures of proteins that are from pathogenic and host organisms and from normal and diseased human tissues.

The PSI collection of structures will assist biomedical investigators in research studies of key biophysical and biochemical problems, such as protein folding, evolution, structure prediction, and the organization of protein families and folds.

Technical developments, the availability of reagents and materials, and experimental outcome data in protein production and crystallization will directly benefit all structural biologists and provide valuable assistance to a broad range of biomedical researchers.

Structural Genomics Centers in US

The Berkeley Structural Genomics Center (BSGC) The BSGC is pursuing an integrated structural genomics program designed to obtain a near-complete structural complement of two minimal genomes, Mycoplasma genitalium and Mycoplasma pneumoniae, two related human and animal pathogens. Both NMR spectroscopy and X-ray crystallography are being used for structural determination.

Center for Eukaryotic Structural Genomics (CESG)
The CESG was founded as a collaborative effort to develop the technologies needed for economical high-throughput structure determination of biologically important eukaryotic proteins and to extend the knowledge of fold-function space. This project also aims to further the research of biologically important proteins in Arabidopsis. The protein structures are being determined via X-ray crystallography or NMR spectroscopy.

The Joint Center for Structural Genomics (JCSG)
The research focus of the JCSG is on the prokaryote *Thermotoga maritima*, and the eukaryote *Caenorhabditis elegans*, and the main proteins of interest are signaling proteins. The goals involve discovering new protein folds, attaining complete coverage of the proteome of the eubacterium *Thermotoga maritima*, and creating a high-throughput system from the point of target selection through structure determination. X-ray crystallography is being used for structural determination.

The Midwest Center for Structural Genomics (MCSG)
The objective of the MCSG is to develop and optimize new, rapid, integrated methods for highly cost-effective determination of protein structures through X-ray crystallography. This project aims to quickly solve a large number of "easy" targets, and in the process develop new, more advanced tools, methods and approaches that can be applied to "unsolved and difficult projects". Protein targets have an emphasis on unknown folds and proteins from disease-causing organisms.

The New York Structural Genomics Research Consortium (NYSGRC)
The NYSGRC aims to develop and use the technology for high-throughput structural and functional studies of proteins from humans and model organisms. The consortium is establishing a fully integrated, high-throughput system for protein family classification and target selection, protein expression, purification, crystallization, and structure determination by X-ray crystallography.

Structural Genomics Centers in US

The Northeast Structural Genomics Consortium (NEGS)
The NEGS is focused on human proteins and proteins from eukaryotic model organisms. The project targets representative proteins to provide "coverage" of fold space, and also proteins that are interesting from a functional genomics perspective. In addition, the center is exploring the complementary aspects of X-ray crystallography and NMR spectroscopy.

The Southeast Collaboratory for Structural Genomics (SECSG)
The objective of the SECSG is to develop and test experimental and computational strategies for high throughput structure determination of proteins by X-ray crystallography and NMR methods and to apply these strategies to scan the entire genome of an organism at a rapid pace. The eukaryotic organisms, *Caenorhabditis elegans*, *Homo sapiens* and an ancestrally-related prokaryotic microorganism having a small genome, *Pyrococcus furiosus*, have been selected as representative genomes.

Structural Genomics of Pathogenic Protozoa Consortium (SGPP)
The SGPP consortium aims to determine and analyze the structures of a large number of proteins from major global pathogenic protozoa including *Leishmania major*, *Trypanosoma brucei*, *Trypanosoma cruzi* and *Plasmodium falciparum*. These organisms are responsible for the diseases: leishmaniasis, sleeping sickness, Chagas' disease and malaria. X-ray crystallography is being used for structural determination.

The TB Structural Genomics Consortium (TB)
The goal of the TB consortium is to determine the structures of over 400 proteins from M. tuberculosis, and to analyze these structures in the context of functional information that currently exists and that is generated by the project. These structures will include about 40 novel folds and 200 new families of protein structures. The protein structures are being determined using X-ray crystallography.

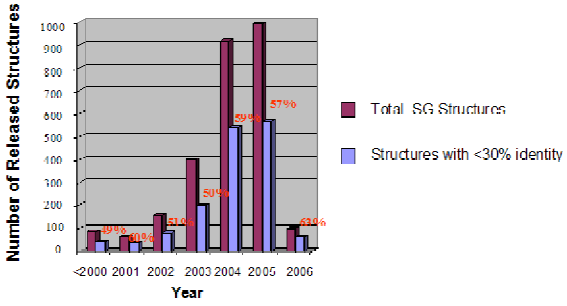
Structural Genomics Progress

Status	Total Number of Targets	(%) Relative to "Cloned" Targets	(%) Relative to "Expressed" Targets	(%) Relative to "Purified" Targets	(%) Relative to "Crystallized" Targets
Cloned	61522	100.00	-	-	-
Expressed	39540	64.27	100.00	-	-
Soluble	18221	29.62	46.08	-	-
Purified	14031	22.81	35.49	100.00	-
Crystallized	5616	9.13	14.20	40.03	100.00
Diffraction-quality Crystals	2909	4.73	7.36	20.73	51.80
Diffraction	2429	3.95	6.14	17.31	43.25
NMR Assigned	1051	1.71	2.66	7.49	-
HSQC	1890	3.07	4.78	13.47	-
Crystal Structure	2291	3.72	5.79	16.33	40.79
NMR Structure	953	1.55	2.41	6.79	-
In PDB	2849	4.63	7.21	20.31	35.15
Work Stopped	14137	-	-	-	-
Test Target	4	-	-	-	-
Other	10	-	-	-	-

Note 1: *Last updated: Mar 24 2006*

~40% of structures are from SG in Europe and Asia

Unique Folds?

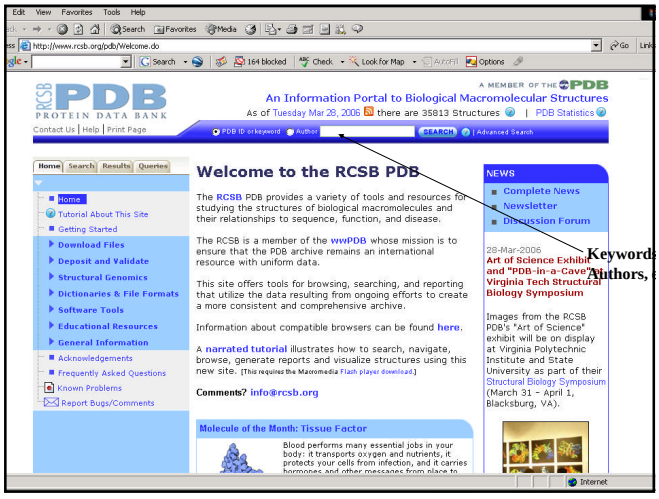


Protein Structure Databases

- **Where does protein structural information reside?**
  - **PDB**:
    - http://www.rcsb.org/pdb/
  - **MMDB**:
    - http://www.ncbi.nlm.nih.gov/Structure/
  - **FSSP**:
    - http://www.ebi.ac.uk/dali/fssp/
  - **SCOP**:
    - http://scop.mrc-lmb.cam.ac.uk/scop/
  - **CATH**:
    - http://www.biochem.ucl.ac.uk/bsm/cath_new/

Jon

Ingo

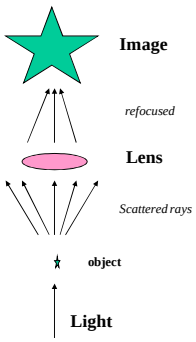


PDB Contents 1 April 2008

		Molecule Type				
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	**Total**
Exp. Method	X-ray	39500	1020	1798	24	**42342**
	NMR	6202	804	137	7	**7150**
	Electron Microscopy	116	11	43	0	**170**
	Other	88	4	4	0	**98**
	Total	**45906**	**1839**	**1982**	**33**	**49760**

X-ray Crystallography

Optical Microscope



Atomic Resolution

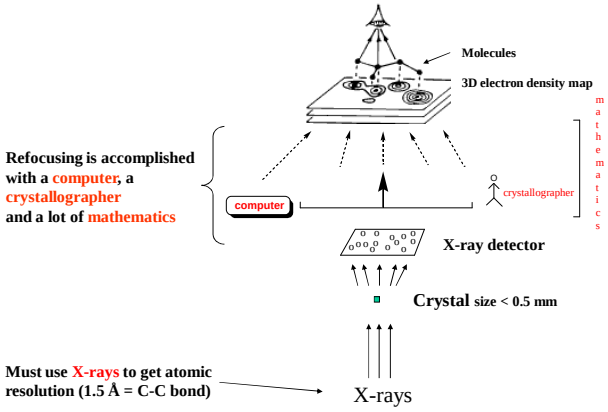
We want to resolve inter-atomic distances (~1.5 Å, 0.15 nM)

Visible light has a wavelength of ~ 500 nm (5000 Å)

Electron beam: λ_e ~ 0.001 Å (if e^- is moving at c)
Electron velocity is less in electron microscopes
Typical resolution is ~10 Å, but can be improved

X-ray generators produce photons of λ = 0.5 – 2.5 Å
Use λ = 1.542 Å

X-ray Crystallography



X-Ray Crystallography

1. Make crystals of your protein
   0.3-1.0mm in size
   Proteins must be in an ordered, repeating pattern.
2. X-ray beam is aimed at crystal and data is collected.
3. Structure is determined from the diffraction data.

X-Ray Crystallography

1. Make crystals of your protein
   0.3-1.0mm in size
   Proteins must be in an ordered, repeating pattern.
2. X-ray beam is aimed at crystal and data is collected.
3. Structure is determined from the diffraction data.

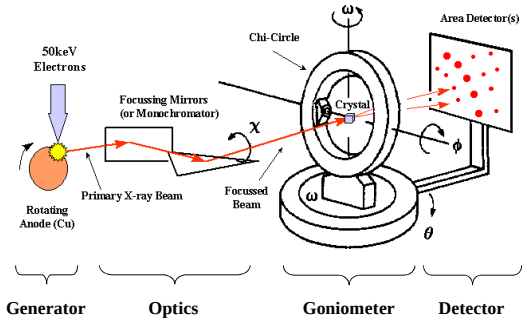
Protein Crystals

Schmid, M. Trends in Microbiolgy, **10**:s27-s31.

X-Ray Crystallography

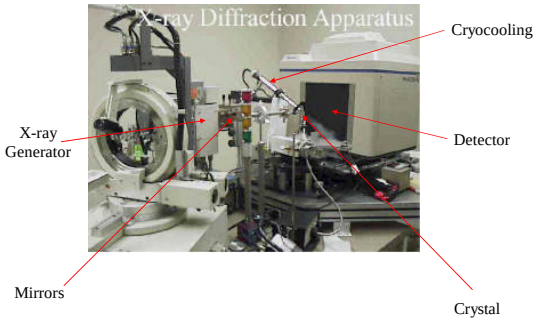
1. Make crystals of your protein
   0.3-1.0mm in size
   Proteins must be in an ordered, repeating pattern.
2. X-ray beam is aimed at crystal and data is collected.
3. Structure is determined from the diffraction data.

X-Ray Diffraction Experiment



Optional: Cryo for protein samples

X-ray Crystallography Equipment



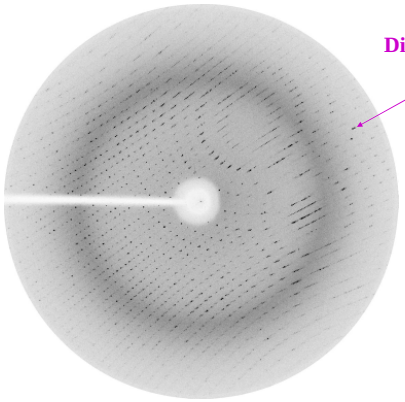
X-Ray Crystallography

1. Make crystals of your protein
   0.3-1.0mm in size
   Proteins must be in an ordered, repeating pattern.

2. X-ray beam is aimed at crystal and data is collected.

3. Structure is determined from the diffraction data.

Protein Diffraction Image



Why Spots?

X-ray scattering from individual proteins is diffuse

Spots arise from a phenomenon called diffraction that is based on the crystal lattice

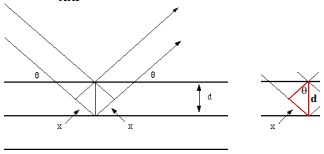
Location of reflections indicates how an object crystallized
230 possibilities

Intensity of reflections contains information about the structure of the object in the crystal

Bragg's Law

Why do we get spots (reflections) and not a diffuse pattern of scattered x-rays?

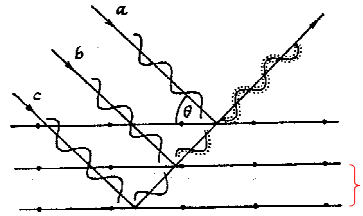
$$2d_{hkl} \sin\theta = n\lambda$$



sin θ = x/d
2d sin θ = 2x

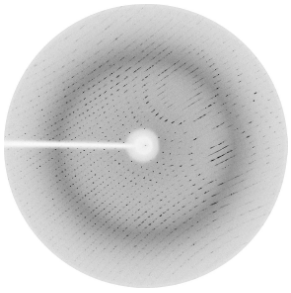
Difference in path (2x) must equal integral number of wavelengths (nλ)

Constructive Interference



- Condition for reflection

Phase Problem



Every diffraction spot (reflection) has a phase and intensity
- **-The intensities are recorded by the detector**
- **-The phases are lost**
- **-Must have both to reconstruct the image (structure)**

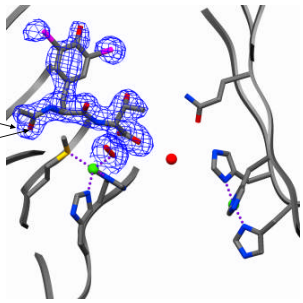
Solutions to the Phase Problem

Molecular replacement
- **-Use known structure of close homologue**
- **-Rotational and translational search for solution**

Heavy atom labeling
- **-Label the protein with electron dense atoms (Hg)**
- **-Compare independent datasets collected from native and labeled protein**
- **-Heavy atom substructure provides initial phases**

Anomalous diffraction
- **-Crystal must contain atoms with absorption edges between 0.5 and 2.5 Å**
- **-Compare independent datasets collected at pre-edge and post-edge x-ray energies**

Model Building



Crystallography Pros/Cons

Advantages

- **-can be ”fast” – down to a few months**
- **-large structures possible (ribosome)**
- **-very low resolution (down to 0.5 Å)**
- **-observables typically > refinement parameters**

Disadvantages

- **-requires crystal formation**
- **-non-physiological conditions**
- **-crystal contacts can limit protein motion**

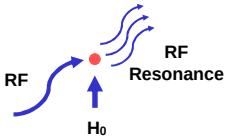
Nuclear Magnetic Resonance

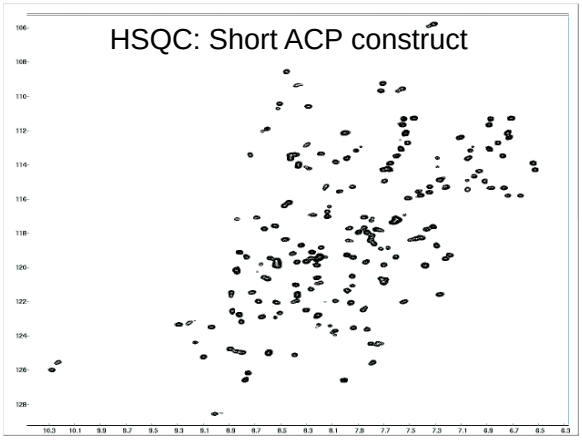
Nuclear Magnetic Resonance

Magnetically align unpaired proton spins (H_0)

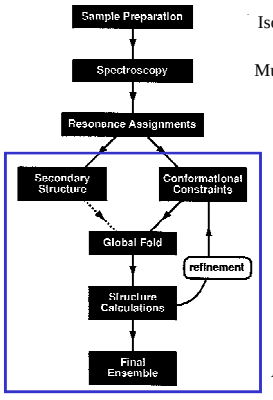
Probe with radio frequency (RF)

Observe resonance





NMR Overview



Isotopic labeling (^{15}N, ^{13}C)

Multiple experiments (pulse sequences)

Positional refinement typically not possible. Dihedral angles used.

As many as 20 structures produced

NMR Experimental Observables

- Backbone conformation from chemical shifts (Chemical Shift Index- CSI)
- Distance constraints from NOEs
- Hydrogen bond constraints
- Backbone and side chain dihedral angle constraints from scalar couplings
- Orientation constraints from residual dipolar couplings

NMR Pros/Cons

Advantages

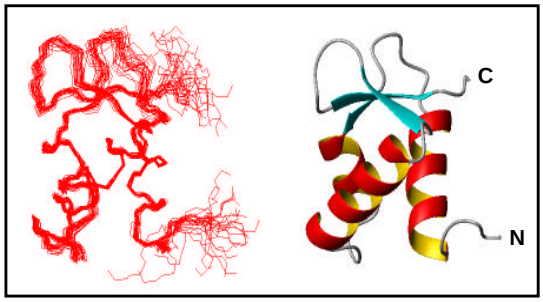
-no crystal formation needed
-more physiological conditions

Disadvantages

-results in a set of models that are compatible with data
-size limitation to 200-300 residues (extended recently)
-must label protein with ^{15}N and ^{13}C
-observables typically < refinement parameters

Precision

NMR vs. X-ray



RMSD of the ensemble Mean coordinate error

A PDB File

Header contains information about protein and structure
date of the entry, references, crystallographic data,
contents and positions of secondary structure elements

```
HEADER    OXIDOREDUCTASE                  03-OCT-02   1MXT
TITLE     ATOMIC RESOLUTION STRUCTURE OF CHOLESTEROL OXIDASE
TITLE    2 (STREPTOMYCES SP. SA-COO)
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CHOLESTEROL OXIDASE;
COMPND   3 CHAIN: A;
COMPND   4 SYNONYM: CHOD;
COMPND   5 EC: 1.1.3.6;
COMPND   6 ENGINEERED: YES;
COMPND   7 OTHER_DETAILS: FAD COFACTOR NON-COVALENTLY BOUND TO THE
COMPND   8 ENZYME
```

A PDB File

Header contains information about protein and structure
date of the entry, references, crystallographic data,
contents and positions of secondary structure elements

```
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: STREPTOMYCES SP.;
SOURCE   3 ORGANISM_COMMON: BACTERIA;
SOURCE   4 GENE: CHOA;
SOURCE   5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   6 EXPRESSION_SYSTEM_COMMON: BACTERIA;
SOURCE   7 EXPRESSION_SYSTEM_STRAIN: BL21(DE3)PLYSS;
SOURCE   8 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE   9 EXPRESSION_SYSTEM_PLASMID: PCO202
```

A PDB File

Header contains information about protein and structure
date of the entry, references, crystallographic data,
contents and positions of secondary structure elements

```
AUTHOR    A.VRIELINK,P.I.LARIO
REVDAT   1   25-FEB-03 1MXT    0
JRNL      AUTH   P.I.LARIO,N.SAMPSON,A.VRIELINK
JRNL      TITL   SUB-ATOMIC RESOLUTION CRYSTAL STRUCTURE OF
JRNL      TITL 2 CHOLESTEROL OXIDASE: WHAT ATOMIC RESOLUTION
JRNL      TITL 3 CRYSTALLOGRAPHY REVEALS ABOUT ENZYME MECHANISM AND
JRNL      TITL 4 THE ROLE OF FAD COFACTOR IN REDOX ACTIVITY
JRNL      REF    J.MOL.BIOL.                V. 326  1635 2003
JRNL      REFN   ASTM JMOBAK  UK ISSN 0022-2836
```

A PDB File

Header contains information about protein and structure
date of the entry, references, crystallographic data,
contents and positions of secondary structure elements

```
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 0.95
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 28.00
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   COMPLETENESS FOR RANGE        (%) : 94.1
REMARK   3   CROSS-VALIDATION METHOD           : FREE R
REMARK   3   FREE R VALUE TEST SET SELECTION   : RANDOM
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT (NO CUTOFF).
REMARK   3   R VALUE   (WORKING + TEST SET, NO CUTOFF) : 0.110
REMARK   3   R VALUE           (WORKING SET, NO CUTOFF) : 0.110
REMARK   3   FREE R VALUE                (NO CUTOFF) : 0.132
REMARK   3   FREE R VALUE TEST SET SIZE (%, NO CUTOFF) : 5.000
REMARK   3   FREE R VALUE TEST SET COUNT  (NO CUTOFF) : 13180

REMARK   3   TOTAL NUMBER OF REFLECTIONS  (NO CUTOFF) : 263551
```

Resolution:
Low > 3 Å
Mid 2-3 Å
High 1.5-2 Å
Very High < 1.5 Å

R factor (residual):
Low resolution ~ 27%
Mid resolution ~ 22 %
High resolution ~ 29 %
Very High res ~ 15%

A PDB File

Header contains information about protein and structure
date of the entry, references, crystallographic data,
contents and positions of secondary structure elements

```
HELIX    14  14 ALA A  289  THR A  304  1                                  16
HELIX    15  15 THR A  402  GLN A  405  5                                   4
HELIX    16  16 ASN A  406  GLY A  425  1                                  20
HELIX    17  17 ASP A  474  ILE A  478  5                                   5
HELIX    18  18 PRO A  486  VAL A  506  1                                  21
SHEET    1   A 6 HIS A 248  GLN A 255  0
SHEET    2   A 6 TYR A 261  LYS A 268 -1  O  GLU A 266   N  GLN A 249
SHEET    3   A 6 LEU A 274  LEU A 287 -1  O  LEU A 275   N  GLN A 267
SHEET    4   A 6 TYR A  10  ILE A  16  1  N  VAL A  14   O  PHE A 286
SHEET    5   A 6 THR A  36  GLU A  40  1  O  LEU A  37   N  VAL A  15

SHEET    6   A 6 VAL A 242  THR A 246  1  O  THR A 243   N  MET A  38
```

A PDB File

Body of PDB file contains information about the atoms in the structure

```
ATOM     76  N   PRO A  12      31.129  -4.659  43.245  1.00  9.00           N
ATOM     77  CA  PRO A  12      32.426  -4.662  42.542  1.00  9.00           C
ATOM     78  C   PRO A  12      32.423  -4.009  41.182  1.00  8.02           C
ATOM     79  O   PRO A  12      33.267  -3.177  40.892  1.00  8.31           O
ATOM     80  CB  PRO A  12      32.791  -6.126  42.592  1.00 10.02           C
ATOM     81  CG  PRO A  12      32.190  -6.663  43.857  1.00 10.12           C
ATOM     82  CD  PRO A  12      30.850  -5.927  43.925  1.00  9.87           C
ATOM     90  N   ALA A  13      31.485  -4.468  40.316  1.00  8.06           N
ATOM     91  CA  ALA A  13      31.357  -3.854  39.004  1.00  7.28           C
ATOM     92  C   ALA A  13      29.947  -3.309  38.814  1.00  7.21           C
ATOM     93  O   ALA A  13      28.969  -3.932  39.200  1.00  7.56           O
ATOM     94  CB  ALA A  13      31.636  -4.879  37.897  1.00  8.54           C
```

Atom number
Residue name
Residue number
Atom name

A PDB File

Body of PDB file contains information about the atoms in the structure

```
ATOM     76  N   PRO A  12      31.129  -4.659  43.245  1.00  9.00           N
ATOM     77  CA  PRO A  12      32.426  -4.662  42.542  1.00  9.00           C
ATOM     78  C   PRO A  12      32.423  -4.009  41.182  1.00  8.02           C
ATOM     79  O   PRO A  12      33.267  -3.177  40.892  1.00  8.31           O
ATOM     80  CB  PRO A  12      32.791  -6.126  42.592  1.00 10.02           C
ATOM     81  CG  PRO A  12      32.190  -6.663  43.857  1.00 10.12           C
ATOM     82  CD  PRO A  12      30.850  -5.927  43.925  1.00  9.87           C
ATOM     90  N   ALA A  13      31.485  -4.468  40.316  1.00  8.06           N
ATOM     91  CA  ALA A  13      31.357  -3.854  39.004  1.00  7.28           C
ATOM     92  C   ALA A  13      29.947  -3.309  38.814  1.00  7.21           C
ATOM     93  O   ALA A  13      28.969  -3.932  39.200  1.00  7.56           O
ATOM     94  CB  ALA A  13      31.636  -4.879  37.897  1.00  8.54           C
```

Coordinates in Å

(X,Y,Z)

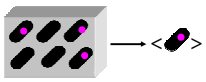
Mean coordinate error:

Low > 3 Å	.4 Å
Mid 2-3 Å	.3 Å
High 1.5-2 Å	.2 Å
Very High < 1.5 Å	.1 Å

A PDB File

Body of PDB file contains information about the atoms in the structure

```
ATOM     76  N   PRO A  12      31.129  -4.659  43.245  1.00  9.00           N
ATOM     77  CA  PRO A  12      32.426  -4.662  42.542  1.00  9.00           C
ATOM     78  C   PRO A  12      32.423  -4.009  41.182  1.00  8.02           C
ATOM     79  O   PRO A  12      33.267  -3.177  40.892  1.00  8.31           O
ATOM     80  CB  PRO A  12      32.791  -6.126  42.592  1.00 10.02           C
ATOM     81  CG  PRO A  12      32.190  -6.663  43.857  1.00 10.12           C
ATOM     82  CD  PRO A  12      30.850  -5.927  43.925  1.00  9.87           C
ATOM     90  N   ALA A  13      31.485  -4.468  40.316  1.00  8.06           N
ATOM     91  CA  ALA A  13      31.357  -3.854  39.004  1.00  7.28           C
ATOM     92  C   ALA A  13      29.947  -3.309  38.814  1.00  7.21           C
ATOM     93  O   ALA A  13      28.969  -3.932  39.200  1.00  7.56           O
ATOM     94  CB  ALA A  13      31.636  -4.879  37.897  1.00  8.54           C
```



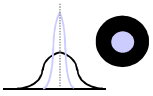
- Occupancy of 0.5

Fractional occupancy

A PDB File

Body of PDB file contains information about the atoms in the structure

```
ATOM     76  N   PRO A  12      31.129  -4.659  43.245  1.00  9.00           N
ATOM     77  CA  PRO A  12      32.426  -4.662  42.542  1.00  9.00           C
ATOM     78  C   PRO A  12      32.423  -4.009  41.182  1.00  8.02           C
ATOM     79  O   PRO A  12      33.267  -3.177  40.892  1.00  8.31           O
ATOM     80  CB  PRO A  12      32.791  -6.126  42.592  1.00 10.02           C
ATOM     81  CG  PRO A  12      32.190  -6.663  43.857  1.00 10.12           C
ATOM     82  CD  PRO A  12      30.850  -5.927  43.925  1.00  9.87           C
ATOM     90  N   ALA A  13      31.485  -4.468  40.316  1.00  8.06           N
ATOM     91  CA  ALA A  13      31.357  -3.854  39.004  1.00  7.28           C
ATOM     92  C   ALA A  13      29.947  -3.309  38.814  1.00  7.21           C
ATOM     93  O   ALA A  13      28.969  -3.932  39.200  1.00  7.56           O
ATOM     94  CB  ALA A  13      31.636  -4.879  37.897  1.00  8.54           C
```



B-factor Å^2

Visualization of Structures