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 ~64,623 protein structures  
  Genebank ~61,132,599 sequences

• Of the 64,623 structures, only 15,702 are dissimilar in sequence (<30% ID).
Growth of GenBank
Now over 100M sequences and 100B base pairs from 223K species

Exponential Increase

Growth of Structural Data
Currently 64,823 structures deposited

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 (20 April 10)

<table>
<thead>
<tr>
<th>Sequence identity</th>
<th>Number of non-redundant chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>25615</td>
</tr>
<tr>
<td>70%</td>
<td>23116</td>
</tr>
<tr>
<td>50%</td>
<td>20306</td>
</tr>
<tr>
<td>30%</td>
<td>15702</td>
</tr>
</tbody>
</table>

---

Unique folds in PDB (SCOP)
Unique topologies in PDB (CATH)

New Topologies and Folds Becoming Rare

Structural Genomics

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.


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.
The Joint Center for Structural Genomics (JCSG)

During PSI-2, the JCSG has contributed to the overall goal of maximizing structural coverage of protein families with no structural representation and has continued to develop and disseminate innovative new technologies for structural biology. The JCSG consortium theme is the “central machinery of life” — proteins that are conserved in all kingdoms of life.

The Midwest Center for Structural Genomics (MCSG)

In PSI-2, the multi-institutional consortium is rapidly determining the structures of large numbers of strategically selected proteins using x-ray crystallography both to provide structural coverage of major protein superfamilies and to elucidate the entire protein folding space.

The New York Structural Genomics Research Consortium (NYSGRC)

During PSI-2, the consortium’s individual project focuses on new targets, principally protein phosphatases and multifunctional eukaryotic proteins.

The Northeast Structural Genomics Consortium (NEGS)

In PSI-2, the consortium is solving both prokaryotic and eukaryotic structural representatives from the major domain families constituting the eukaryotic proteome.

The 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 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.

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 organism, Caenorhabditis elegans, is being used as an example related prokaryotic microorganism having a small genome, Pyrococcus furiosus, have been selected as representative genomes.

The 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 80 novel folds and 350 new families of protein structures. The protein structures are being determined using X-ray crystallography.
Current PSI Centers

**Large-Scale Centers**
- Joint Center for Structural Genomics
- Midwest Center for Structural Genomics
- New York SGX Research Center for Structural Genomics
- Northeast Structural Genomics Consortium

**Specialized Centers**
- Accelerated Technologies Center for Gene to 3D Structure
- Center for High-Throughput Structural Biology
- Center for Structures of Membrane Proteins
- Integrated Center for Structure and Function Innovation
- New York Consortium on Membrane Protein Structures

**Homology Modeling Centers**
- Joint Center for Molecular Modeling
- New Methods for High-Resolution Comparative Modeling

**Resource Centers**
- PSI-Materials Repository
- PSI Knowledgebase

---

### 2008 Structural Genomics Progress

<table>
<thead>
<tr>
<th>Status</th>
<th>Total Number of Targets (%)</th>
<th>(% Relative to &quot;Cloned&quot; Targets)</th>
<th>(% Relative to &quot;Expressed&quot; Targets)</th>
<th>(% Relative to &quot;Purified&quot; Targets)</th>
<th>(% Relative to &quot;Crystallized&quot; Targets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed</td>
<td>1212</td>
<td>100.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Expressed</td>
<td>1058</td>
<td>87.2</td>
<td>100.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Soluble</td>
<td>584</td>
<td>48.1</td>
<td>50.6</td>
<td>100.0</td>
<td>-</td>
</tr>
<tr>
<td>Purified</td>
<td>298</td>
<td>24.5</td>
<td>26.8</td>
<td>34.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Diffraction</td>
<td>190</td>
<td>15.7</td>
<td>16.5</td>
<td>21.2</td>
<td>50.5</td>
</tr>
<tr>
<td>Diffraction</td>
<td>7</td>
<td>10.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NMR Assigned</td>
<td>14</td>
<td>1.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Crystal Structure</td>
<td>20</td>
<td>1.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>In PDB</td>
<td>12</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Test Target</td>
<td>21</td>
<td>1.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other</td>
<td>23</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

---

### 2010 Structural Genomics Progress

<table>
<thead>
<tr>
<th>Status</th>
<th>Total Number of Targets (%)</th>
<th>(% Relative to &quot;Cloned&quot; Targets)</th>
<th>(% Relative to &quot;Expressed&quot; Targets)</th>
<th>(% Relative to &quot;Purified&quot; Targets)</th>
<th>(% Relative to &quot;Crystallized&quot; Targets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed</td>
<td>1026</td>
<td>100.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Expressed</td>
<td>895</td>
<td>87.4</td>
<td>100.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Soluble</td>
<td>540</td>
<td>53.2</td>
<td>56.5</td>
<td>100.0</td>
<td>-</td>
</tr>
<tr>
<td>Purified</td>
<td>286</td>
<td>27.8</td>
<td>29.9</td>
<td>35.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Diffraction</td>
<td>155</td>
<td>15.2</td>
<td>16.4</td>
<td>21.1</td>
<td>50.4</td>
</tr>
<tr>
<td>Diffraction</td>
<td>5</td>
<td>5.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NMR Assigned</td>
<td>25</td>
<td>2.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Crystal Structure</td>
<td>11</td>
<td>1.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>In PDB</td>
<td>10</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Test Target</td>
<td>22</td>
<td>2.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other</td>
<td>21</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

~36% of structures are from SG in Europe and Asia
Protein Structure Databases

- **Where does protein structural information reside?**
  - **PDB:**
    - [http://www.rcsb.org/pdb/](http://www.rcsb.org/pdb/)
  - **MMDB:**
  - **FSSP:**
    - [http://www.ebi.ac.uk/dali/hsp/](http://www.ebi.ac.uk/dali/hsp/)
  - **SCOP:**
    - [http://scop.mrc-lmb.cam.ac.uk scop/](http://scop.mrc-lmb.cam.ac.uk scop/)
  - **CATH:**
    - [http://www.biochem.ucl.ac.uk/bsm/cath_new/](http://www.biochem.ucl.ac.uk/bsm/cath_new/)
PDB Contents 20 April 2010

<table>
<thead>
<tr>
<th>Exp. Method</th>
<th>Proteins</th>
<th>Nucleic Acids</th>
<th>Protein/NA Complexes</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-RAY</td>
<td>52212</td>
<td>1206</td>
<td>2401</td>
<td>17</td>
<td>55836</td>
</tr>
<tr>
<td>NMR</td>
<td>729</td>
<td>896</td>
<td>154</td>
<td>7</td>
<td>8336</td>
</tr>
<tr>
<td>ELECTRON MICROSCOPY</td>
<td>195</td>
<td>17</td>
<td>76</td>
<td>0</td>
<td>288</td>
</tr>
<tr>
<td>HYBRID</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>Other</td>
<td>123</td>
<td>4</td>
<td>4</td>
<td>13</td>
<td>144</td>
</tr>
<tr>
<td>Total</td>
<td>59825</td>
<td>2124</td>
<td>2636</td>
<td>38</td>
<td>64623</td>
</tr>
</tbody>
</table>

X-ray Crystallography
Visible light has a wavelength of ~500 nm (5000 Å)

Electron beam: $\lambda_e \approx 0.001 \text{ Å}$ (if $c$ 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 $\lambda = 0.5 – 2.5 \text{ Å}$
   - Use $\lambda = 1.542 \text{ Å}$
   - Refocusing is accomplished with a computer, a crystallographer and a lot of mathematics
   - Must use X-rays to get atomic resolution ($1.5 \text{ Å} = \text{C-C bond}$)
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

1. Make crystals of your protein
   0.3-1.0 mm 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

Generator Optics Goniometer Detector

Optional: Cryo for protein samples

X-ray Crystallography Equipment

X-ray Generator

Moirrors

Crystal

Cryocooling

Detector

X-ray Diffraction Apparatus

Video-microscope

Cryocooler

CCD detector

X-ray optics
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?

\[ 2d_{nkl} \sin \theta = n\lambda \]

\[ \sin \theta = \frac{x}{d} \]

\[ 2d \sin \theta = 2x \]

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

Constructive Interference

• Condition for reflection

Resolution

Accuracy and detail

https://www.ruppweb.org/Garland/gallery/Ch1/index.htm#5
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

Electron density (data)
Structure (model)

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

Magnetically align unpaired proton spins ($H_0$)

Probe with radio frequency (RF)

Observe resonance

HSQC: Short ACP construct
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

A PDB File

Header contains information about protein and structure
date of the entry, references, crystallographic data,
contents and positions of secondary structure elements
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
REVDATE   1MXT    0
JRNL        AUTH   P.I.LARIO,N.SAMPSON,A.VRIELINK
JRNL        TITL   SUB-A TOMIC RESOLUTION CRYSTAL STRUCTURE OF
JRNL        TITL 2 CHOLESTEROL OXIDASE: WHA T A TOMIC 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

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

Coordinates in Å

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

<table>
<thead>
<tr>
<th>ATOM</th>
<th>SEQ</th>
<th>NAME</th>
<th>RESID</th>
<th>ALT</th>
<th>LABEL</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>TEMP</th>
<th>OCCUP</th>
<th>B-FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM1</td>
<td>76</td>
<td>N</td>
<td>PRO</td>
<td>A</td>
<td>12</td>
<td>31.129</td>
<td>-4.659</td>
<td>43.245</td>
<td>1.00</td>
<td>9.00</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM2</td>
<td>77</td>
<td>CA</td>
<td>PRO</td>
<td>A</td>
<td>12</td>
<td>32.426</td>
<td>-4.662</td>
<td>42.542</td>
<td>1.00</td>
<td>9.00</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM3</td>
<td>78</td>
<td>C</td>
<td>PRO</td>
<td>A</td>
<td>12</td>
<td>32.423</td>
<td>-4.009</td>
<td>41.182</td>
<td>1.00</td>
<td>8.02</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM4</td>
<td>79</td>
<td>O</td>
<td>PRO</td>
<td>A</td>
<td>12</td>
<td>33.267</td>
<td>-3.177</td>
<td>40.892</td>
<td>1.00</td>
<td>8.31</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM5</td>
<td>80</td>
<td>CB</td>
<td>PRO</td>
<td>A</td>
<td>12</td>
<td>32.791</td>
<td>-6.126</td>
<td>42.592</td>
<td>1.00</td>
<td>10.02</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM6</td>
<td>81</td>
<td>CG</td>
<td>PRO</td>
<td>A</td>
<td>12</td>
<td>32.190</td>
<td>-6.663</td>
<td>43.857</td>
<td>1.00</td>
<td>10.12</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM7</td>
<td>82</td>
<td>CD</td>
<td>PRO</td>
<td>A</td>
<td>12</td>
<td>30.850</td>
<td>-5.927</td>
<td>43.925</td>
<td>1.00</td>
<td>9.87</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM8</td>
<td>90</td>
<td>N</td>
<td>ALA</td>
<td>A</td>
<td>13</td>
<td>31.485</td>
<td>-4.468</td>
<td>40.316</td>
<td>1.00</td>
<td>8.06</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM9</td>
<td>91</td>
<td>CA</td>
<td>ALA</td>
<td>A</td>
<td>13</td>
<td>31.357</td>
<td>-3.854</td>
<td>39.004</td>
<td>1.00</td>
<td>7.28</td>
<td>0.000</td>
</tr>
<tr>
<td>ATOM10</td>
<td>92</td>
<td>C</td>
<td>ALA</td>
<td>A</td>
<td>13</td>
<td>29.969</td>
<td>-4.855</td>
<td>38.814</td>
<td>1.00</td>
<td>7.21</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Fractional occupancy

Occupancy of 0.5

B-factor Å²