Protein Structure Determination

How are these structures determined?

Protein Sequences Far Outnumber Structures

- Only a small number of protein structures have been experimentally determined.
  - PDB ~30,500 protein structures
  - Genebank ~42,000,000 sequences
- Of the 30,500 structures, only about 7000 are unique.

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.

Growth of Structural Data

Currently 30,453 structures deposited

Growth of GenBank

Release 146 (Feb 2005) has 46,849,831,226 base pairs

Exponential Increase

Growth of Structural Data

Weak Exponential?
Structural Proteomics

- Use experimentally determined structures to model the structures of similar proteins
  - Threading
  - Homology Modeling
  - Fold recognition
- 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
- Avoids Ab initio structure determination

Redundancy in PDB (19 April 05)

<table>
<thead>
<tr>
<th>Sequence identity</th>
<th>Number of non-redundant chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>10503</td>
</tr>
<tr>
<td>70%</td>
<td>9361</td>
</tr>
<tr>
<td>50%</td>
<td>8009</td>
</tr>
<tr>
<td>30%</td>
<td>6120</td>
</tr>
</tbody>
</table>

Unique folds in PDB

New Folds Becoming Rare

Why?

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.

Structural Genomics Centers

The Berkeley Structural Genomics Center (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 BSGC 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 proteins structures are being determined via X-ray crystallography or NMR spectroscopy.

The research focus of the JCSG is on the prokaryote Thermotoga maritima, and the eukaryotic Caenorhabditis elegans, and the main proteins of interest are signaling proteins. The goal is to discover new protein folds, obtaining complete coverage of the proteome of the archaebacterium 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.

Ingo

Structural Genomics Centers

The Northeast Structural Genomics Consortium (NEGS)

The NEGS is focused on human proteins and proteins from eukaryotic model organisms. The project targets 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 objective of the NEGS 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 archaeology-related prokaryotic microorganism having a small genome, Pyrococcus furiosus, have been selected as representative genomes.

The SGCC consortium aims to develope and analyze the structures of a large number of proteins from major global pathogenic proteins 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 goal of the TSB consortium is to determine the structures of over 400 proteins from M. tuberculosis, and to analyze those 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.

Protein Structure Databases

- **Where does protein structural information reside?**
  - **PDB:**
    - http://www.rcsb.org/pdb/
  - **MMDB:**
  - **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/

PDB Contents 19 April 2005

<table>
<thead>
<tr>
<th>Molecule Type</th>
<th>Proteins, Peptides, and Viruses</th>
<th>Protein/Sugar Complexes</th>
<th>Nucleic Acids</th>
<th>Carbohydrates</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. Tech.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X-ray Diffraction and other</td>
<td>24015</td>
<td>1151</td>
<td>781</td>
<td>11</td>
<td>25956</td>
</tr>
<tr>
<td>NMR</td>
<td>3733</td>
<td>111</td>
<td>649</td>
<td>2</td>
<td>4490</td>
</tr>
<tr>
<td>Total</td>
<td>27748</td>
<td>1262</td>
<td>1430</td>
<td>13</td>
<td>50453</td>
</tr>
</tbody>
</table>

X-ray Crystallography

Keywords: Authors, etc.
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: \( \lambda \approx 0.001 \text{ Å} \) (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 \( \lambda = 0.5 – 2.5 \text{ Å} \)
Use \( \lambda = 1.542 \text{ Å} \)

X-ray Crystallography

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

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

4. 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.

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

4. Structure is determined from the diffraction data.

Protein Crystals

X-Ray Crystallography

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

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

4. Structure is determined from the diffraction data.

X-Ray Crystallography Equipment

X-Ray Diffraction Experiment

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

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

4. Structure is determined from the diffraction data.

Why Spots?

X-ray diffraction from individual proteins is diffuse

Spots arise due to 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?

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

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

\[ \sin \theta = \frac{x}{d}, \quad 2d \sin \theta = 2x \]

Difference in path (2x) must equal integral number of wavelengths (n\lambda)

Constructive Interference

• Condition for reflection

\[ d = \text{resolution} \]

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