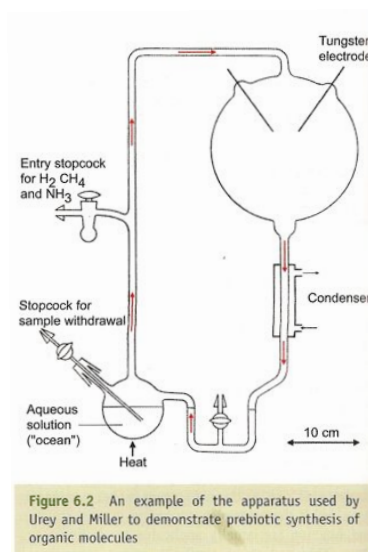
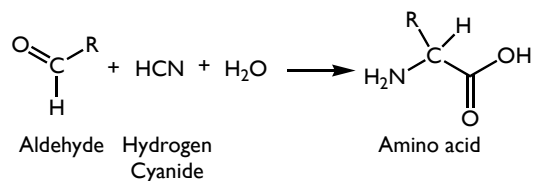


# Protein Diversity

- All organisms are based on proteins assembled using the same 20 amino acid lexicon.
- Principles defining protein structures and functions are therefore applicable to all living systems.
- All living systems are evolutionarily related sharing primitive ancestors that at some point in evolution settled on proteins assembled from the basic 20 amino acids.

# Prebiotic Synthesis and the Origins of Proteins

- In a prebiotic world (~3.6 billion years ago), the building blocks of life (amino acids, nucleotides and carbohydrates) must have arisen via non-biological processes.
- Stanley Miller and Harold Urey simulated primitive conditions consistent with those on Earth ~4 billion years ago and demonstrated that these conditions could result in the production of biomolecules.
  - ◆ Water and a reducing atmosphere containing CO<sub>2</sub>, NH<sub>3</sub> and H<sub>2</sub>S.
  - ◆ Heat, UV and electric pulses.
- Thermal vents on the ocean floor release hot gases and minerals into the ocean. Organic synthesis similar to that observed by Miller and Urey occurs at these vents.



*Proteins: Structure and Function* by David Whitford, 2005, p162

# Prebiotic Synthesis and the Origins of Proteins

Table 6.1 Yields of biomolecules from simulating prebiotic conditions using a mixture of methane, ammonia, water and hydrogen

Biomolecule	Approximate yield (%)
Formic acid	4.0
Glycine	2.1
Glycolic acid	1.9
Alanine	1.7
Lactic acid	1.6
β-Alanine	0.76
Propionic acid	0.66
Acetic acid	0.51
Iminodiacetic acid	0.37
α-Hydroxybutyric acid	0.34
Succinic acid	0.27
Sarcosine	0.25
Iminoaceticpropionic acid	0.13
N-Methylalanine	0.07
Glutamic acid	0.051
N-Methylurea	0.051
Urea	0.034
Aspartic acid	0.024
α-Aminoisobutyric acid	0.007

Shown in red are constituents of proteins (after Miller, S.J. & Orgel, L.E. *The Origins of Life on Earth*. Prentice-Hall, 1975).

*Proteins: Structure and Function* by David Whitford, 2005, p162

# Prebiotic Synthesis and the Origins of Proteins

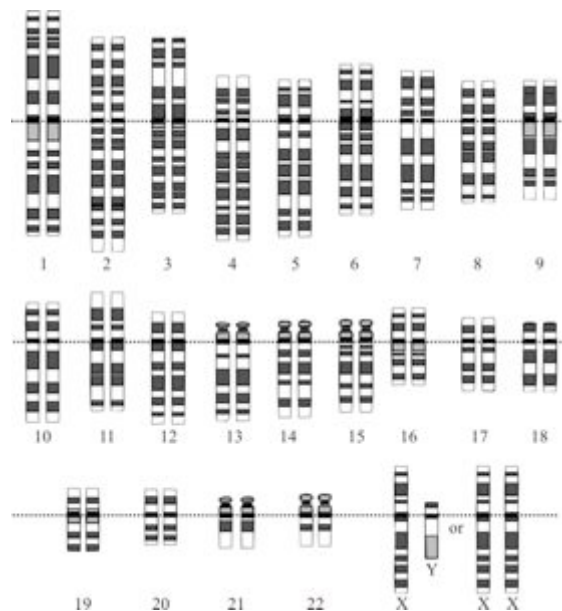
- It is generally believed that the first molecular replicators and “enzymes” were based on RNA.
  - RNA can template its own replication
  - RNA can form complex three-dimensional structures.
  - RNA can function as a catalyst (ribozymes, ribosomes and RNaseP from *E. coli*)
  - Molecules containing adenine units that found in RNA are pervasive (such as NAD, ATP/ADP and FAD).
  - RNA is relatively unstable.
- Over time, these RNA molecules led to the emergence of proteins and DNA... ultimately giving us the current arrangement used by all living organisms\*:

DNA → RNA → Proteins

## Sequence Homology

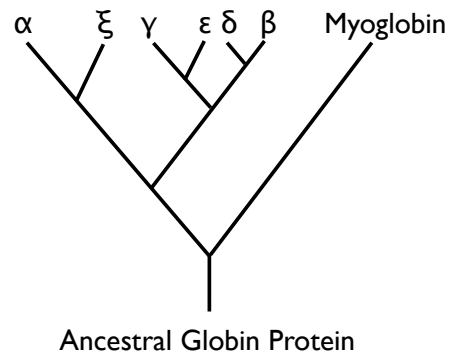
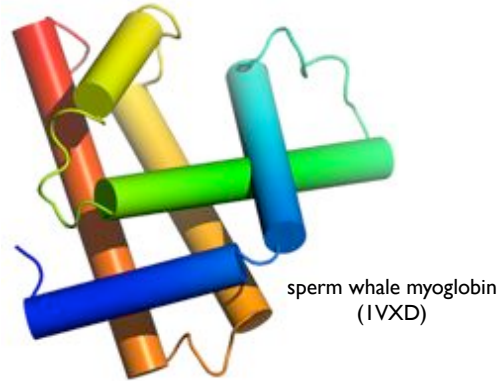
- Recent advances in protein and DNA sequencing (on the genomic scale) have generated massive volumes of data.
- *(it is believed that the human genome contains genes coding for >25,000 polypeptides)*
- Protein and DNA sequence data can be analyzed and compared in order to identify/discern sequence similarities and patterns.
- For any amino acid or nucleotide sequence, a massive number of permutations are possible and similarities do not arise by chance, but instead may reflect evolutionary relationships.
- ♦ **When sequences are evolutionarily linked, the term homology is used to refer to sequence similarities.**

Protein sequences can be similar without being evolutionarily related.



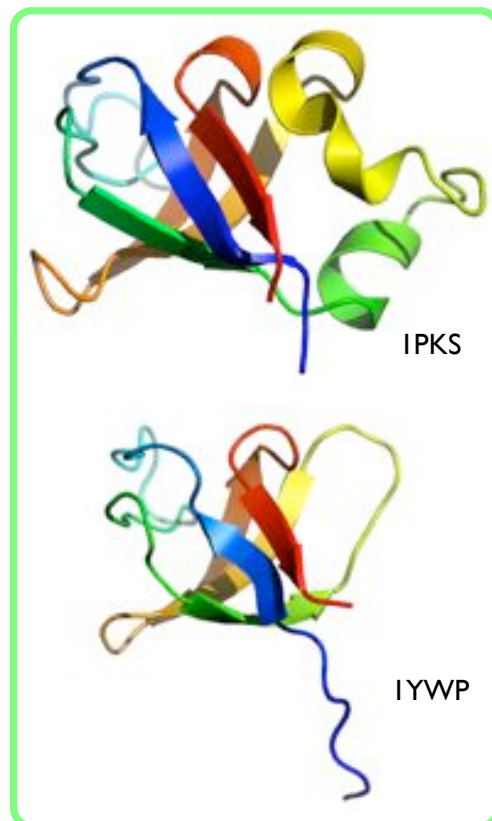
# Gene Fusion and Duplication

- Proteins sharing homologous domains arise as a result of gene duplication.
- Duplicate gene can be subjected to large genetic variation without impacting primary function of the original gene.
- Exemplified by globin family of proteins:
  - Myoglobin monomeric protein likely resembling ancestral protein associated with oxygen storage.
  - Hemoglobin (a tetrameric protein) consists of 2 $\alpha$  and 2 $\beta$  subunits, each of which shows homology with myoglobin.
  - Other globin chains arise during embryogenesis ( $\xi$  and  $\epsilon$ ).
  - Fetal hemoglobin consists of 2 $\alpha$  and 2 $\gamma$  globin subunits.
- Gene duplication coupled with gene fusion results in proteins that are otherwise dissimilar containing homologous domains.



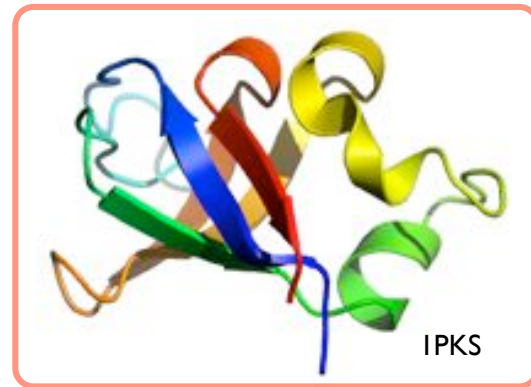
## Sequence Homology

- In order to determine homology it is necessary to establish rules governing potential similarity.
- Alignment of sequences is the first step in determining similarity between two or more sequences.
- Point mutations and larger mutational events occur that give rise to proteins containing different residues, which can obscure the relationships between proteins.
- Comparison to previously characterized proteins aids in the identification newly determined sequences (often from genomic data).
- Domains are key modular elements in proteins, and sequence alignments reveal that gene duplication leads to a proliferation of related domains in different proteins.
- Proteins can be related by the presence of similar domains - SH3 domain is a good example.



# Sequence Homology

- Proteins can be related by the presence of similar domains - SH3 (Src Homology 3) domain is a good example.
- SH3 domains mediate protein-protein interactions - binding to Pro-rich peptide sequences.
- Present in a diverse range of proteins that have little else in common.
- Found in kinases, lipases, GTPases, structural proteins and regulatory proteins:
  - PI3 Kinase
  - CDC24 and CDC25
  - Ras GTPase activating protein
  - Phospholipase
  - Vav proto-oncogene
  - ZAP70
  - GRB2
  - ... and others



# Sequence Homology

- Alignment methods provide a means of graphically representing similarities between query sequences and a library of “known” sequences.
- Numerous alignment strategies, algorithms and scoring systems have been developed.
  - “Pairwise similarity”: comparing each sequence in a database with a query sequence in order to identify matches and sequence similarity.
  - Comparing families of sequences with libraries in order to establish relationships between the sequences.
  - Use known motifs to search database/library for sequences corresponding to the motif. Many protein families have conserved sequence motifs.
- The best sequence alignment methods involve computational methods known as “dynamic programming” that detect optimal pairwise alignment between two or more protein sequences.
- Demands a feasible scoring scheme that reflects degree of relatedness.

National Center for Biotechnology Information (NCBI) was established in 1988 in order to provide a national resource for molecular biology information.

- Creates and maintains databases
- Develops tools for genome data analysis (BLAST)

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

The Expert Protein Analysis System (ExpPASy) was established ~15 years ago. Provides a resource for analyzing protein and genomic data.

- Links to genomic and proteomic databases.
- Tools for analyzing DNA, RNA and protein data.

<http://expasy.org/tools/>

# Sequence Homology

- In the 1970s, Margaret Dayhoff, measured the frequencies with which residues changed as a result of mutation during evolution by carefully aligning the sequences of all known proteins within a single family. She repeated the process for different protein families.
  - Constructed phylogenetic trees for groups of proteins.
  - Yielded a table of relative frequencies describing the rate of residue replacement over an evolutionary period.
- PAM (Point Accepted Mutations) matrices were developed by correlating relative mutation frequencies in closely related proteins and the relative frequency of occurrence of residues in proteins.
  - PAM matrices are effective at scoring similarities in sequences that diverged with evolution.

# Sequence Homology

- An alternative approach, the BLOCKS database, utilizes blocks of ungapped multiple sequence alignments that correspond to conserved regions in proteins. Conserved segments or “blocks” constructed from databases of families of related proteins (databases such as Pfam, ProDom, InterPro or Prosite).
  - ~9,000 “blocks” from ~2,000 protein families.
  - The BLOCKS database has been used to generate substitution matrices that are a fundamental component of prominent alignment programs such as BLAST and FASTA. (used for scoring alignments based on a large set of ~2,000 “blocks”)
  - Matrices based on BLOCKS database can be used to detect distant relationships.
  - In general if query sequence shares 25-30% identity with sequences in the database it likely represents a homologous protein.

**Pfam:** <http://pfam.sanger.ac.uk/>

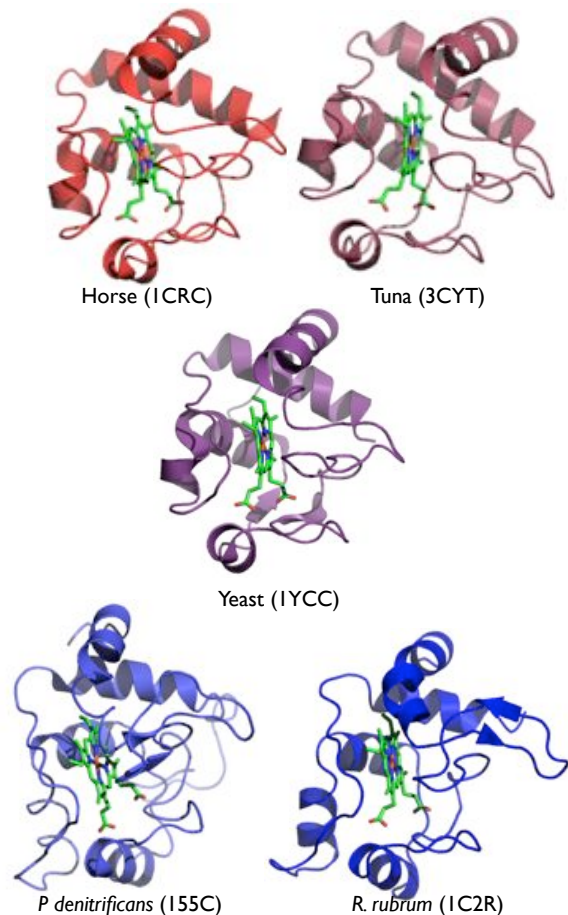
**ProDom:** <http://prodom.prabi.fr/prodom/current/html/home.php>

**InterPro:** <http://www.ebi.ac.uk/interpro/>

**Prosite:** <http://www.expasy.ch/prosite/>

# Structural Homology

- The diversity of folded conformations of proteins is less than would be expected based solely on the potential sequence diversity.
- Protein folds have been conserved throughout evolution despite changes in primary sequence.
- In most cases, structural similarities arise from sequence homology, but in some cases structural homology has been observed even though the evolutionary link is not clear.
  - Cytochrome C family of proteins is a good example of sequence and structural homology (see figure).
  - Only 18 residues conserved between cytochrome C from Horse, Tuna and Yeast mitochondria and cytochrome C<sub>2</sub> of *R. rubrum* and cytochrome C-550 of *P. denitrificans*.



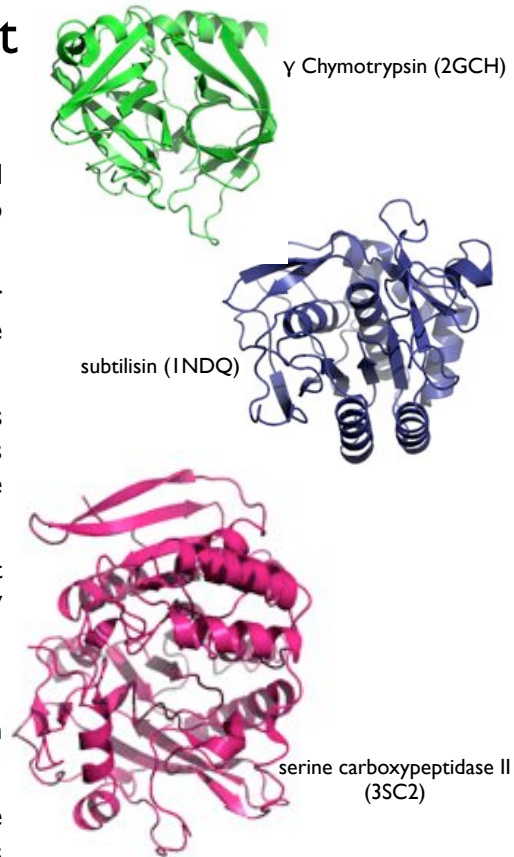
# Structural Homology

- Changes in sequence for different proteins in a family can be used to calculate an average mutation rate.
- “House-keeping” proteins such as histones, essential enzymes and cytoskeletal proteins evolve at very slow rates (usually 1-10 mutations per 100 res per 100 million years).
- “House-keeping” proteins are excellent for tracing evolutionary relationships over hundreds of millions of years.
- Proteins with different structures and functions evolve at significantly different rates. Therefore it is important to use families of proteins to extrapolate rates of protein evolution.

	<i>H. sapiens</i> vs. <i>M. mulatta</i>
Cytochrome C	>90% identity
hemoglobin $\alpha$ and $\beta$ chains	95-97% identity
fibrinopeptides	<70% identity

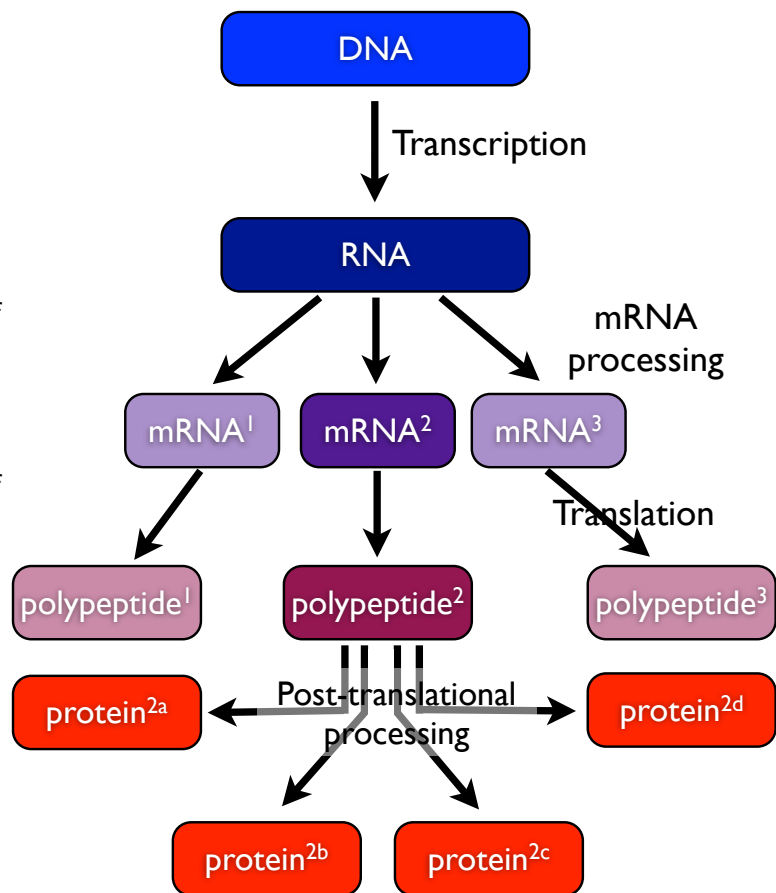
# Structural Homology without Clear Evolutionary Link

- Occasionally, structural homology is observed where there is no clear evolutionary relationship between the proteins (convergent evolution).
- Convergent evolution arises from the use of similar structural motifs and folded conformations in the absence of significant sequence homology.
- A good example of convergent evolution is provided by looking at serine proteases (such as trypsin or chymotrypsin) and subtilisin or serine carboxypeptidase II.
  - Utilize a similar catalytic triad of Ser-His-Asp, but the arrangement of residues in the primary sequences are different.
  - They differ in overall folded conformations as well.
- Unlikely that these proteins share a common ancestor. Therefore result of convergent evolution.
- Other examples include Rossman fold in nucleotide binding domains and  $\beta$ -barrel proteins such as triose phosphate isomerase.



# -omics

- Genomics: (complex)
  - Structural Genomics: analysis of genomic DNA and generation of high-resolution genetic maps for organisms.
  - Functional Genomics: study of gene expression and products of gene expression.
- Proteomics: (more complex)
  - Systematic and global analysis of proteins encoded by genome.
  - Understanding structure and function of proteins.
  - Production and processing



# Protein Databases

- Proteomics is still in its early stages and scientific research is already struggling to develop methods to handle the vast amounts of data being generated on top of the information from genomics (enter - bioinformatics).
- Protein databases have been established in an attempt to organize and categorize protein structural and sequence information. (archived at Protein Data Bank)  
<http://www.rcsb.org/pdb/home/home.do>
- Such databases organize data in hierarchical arrangement -- important in analysis of evolutionary relationships.

SCOP (Structural Classification of Proteins) database contains all known protein structures - sorted according to folding pattern - composition and distribution of 2° structure.

- Primarily manual classification and somewhat subjective.
- Organization hierarchy involves classes of folds, superfamilies, families and domains.

<http://scop.mrc-lmb.cam.ac.uk/scop/>  
<http://en.wikipedia.org/wiki/Scop>

# CATH

<http://www.cathdb.info/>

- CATH database attempts to classify protein folds based on hierarchal system:
  - **Class:** secondary structure composition and packing within a protein (four major classes -  $\alpha$ ,  $\beta$ ,  $\alpha/\beta$  and  $\alpha+\beta$ ).
  - **Architecture:** describes overall shape of the domain - orientations of secondary structure elements (such as  $\beta$  barrel and  $\beta$ - $\alpha$ - $\beta$  sandwich) - ignores connectivity of secondary structure elements. [done manually]
  - **Topology (Folds):** assigned based on overall shape and connectivity of secondary structures.
  - **Homology:** brings together evolutionarily related protein domains. Similarities identified based on sequence then structural comparisons.
- Classification involves both automatic and manual methods.

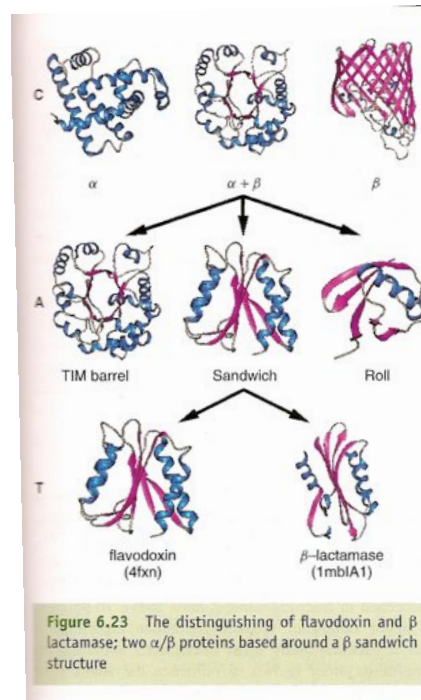


Figure 6.23 The distinguishing of flavodoxin and  $\beta$  lactamase; two  $\alpha/\beta$  proteins based around a  $\beta$  sandwich structure

# Determination of Protein Structures

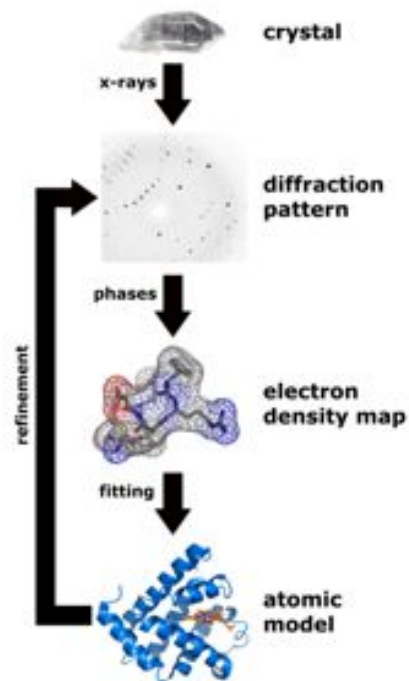
Introduction to Protein Structure - Chapter 18

## Elucidating Protein Structure

- Several different techniques are used to study the structural properties of proteins and other biomacromolecules.
- Protein/peptide primary structure can be determined using chemical and MS methods, or it can be derived from the corresponding gene or cDNA sequence.
- Spectroscopic techniques such as circular dichroism (CD), fluorescence and infrared (IR) can be used to provide insights into structure.
- The main methods for elucidating the three-dimensional structure of proteins and other biomacromolecules are:
  - X-ray crystallography
  - Nuclear magnetic resonance (NMR)

# X-ray Crystallography

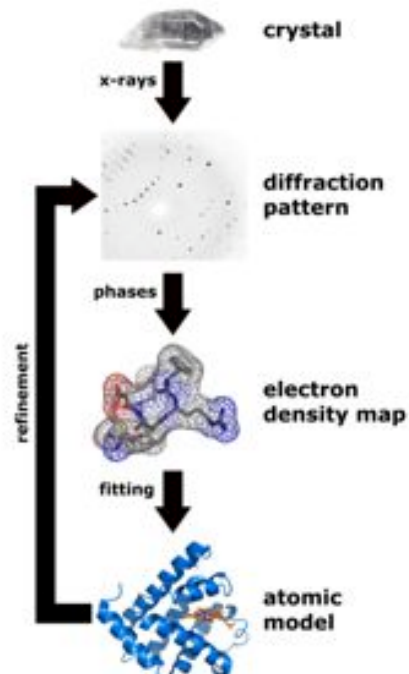
- Solving the three-dimensional structure of a protein by x-ray crystallography requires a well-ordered crystal that strongly diffracts x-rays.
- X-rays are diffracted when they collide with atoms (electrons) of molecules in a crystal.
- A beam of x-rays (monochromatic or polychromatic) is directed at a protein crystal such that the repeating patterning of protein molecules (atoms) in the crystal results in a characteristic diffraction pattern. (Repeating units are called “unit cells” and may contain one or more protein molecules.)
- This requires a pure and homogeneous protein sample (ideally >97% purity).
- X-ray sources:
  - Monochromatic: rotating anode x-ray generators
  - Polychromatic: synchrotrons can generate polychromatic x-ray beams (0.2-2.0 Å) with much greater intensity than achieved by other method.



Source: Wikipedia @ [http://en.wikipedia.org/wiki/Image:X\\_ray\\_diffraction.png](http://en.wikipedia.org/wiki/Image:X_ray_diffraction.png)

# X-ray Crystallography

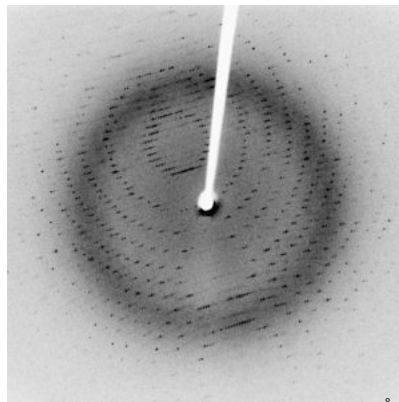
- Crystallization conditions can vary significantly from one protein to another, and it is necessary to determine crystallization conditions for each protein.
  - Slow process (~0.5 mm crystals required)
  - Super saturated solution
  - Hanging-drop method
  - Buffer containing salts and other additives
- Small changes in crystallization conditions (pH, temp., protein conc., salts and other additives) can impact crystal packing and can result in different crystal forms.
- In general, closer packing of protein molecules results better ordering of the molecules in the crystal and a superior diffraction pattern.
- Crystals are usually cooled during exposure to a narrow and parallel beam of x-rays in order to reduce rate of damage by the beam.



Source: Wikipedia @ [http://en.wikipedia.org/wiki/Image:X\\_ray\\_diffraction.png](http://en.wikipedia.org/wiki/Image:X_ray_diffraction.png)

# X-ray Diffraction and Patterns

- A beam of x-rays directed at a protein crystal such that the repeating patterning of protein molecules (atoms) in the crystal results in a characteristic diffraction pattern.
- Most of the x-rays in the beam pass through the crystal, but scattering occurs when x-rays interact with the electrons of atoms in the crystal.
- Regular arrangement of molecules (and their atoms) in the crystals results in interference between the “scattered” x-rays emitted by oscillating electrons (most time canceling out but some add together).
- Diffraction data is usually recorded on an image plate (image is scanned into computer) or by an electron detector (feeding data directly to computer).
- Position of spots on the diffraction pattern can be used to determine size of the unit cell.
- Each diffracted beam (spot in pattern) is defined by three properties:
  - Amplitude: related to spot intensity.
  - Wavelength: determined by x-ray source.
  - Phase: cannot be determined from single experiment.
- Phase problem resolved by:
  - Multiple isomorphous replacement (MIR)
  - Multiwavelength anomalous diffraction (MAD)

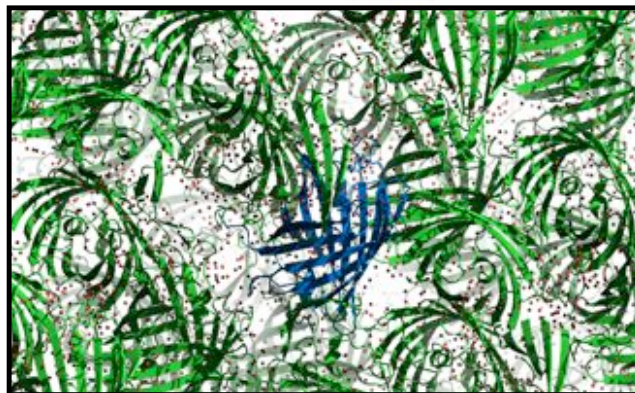
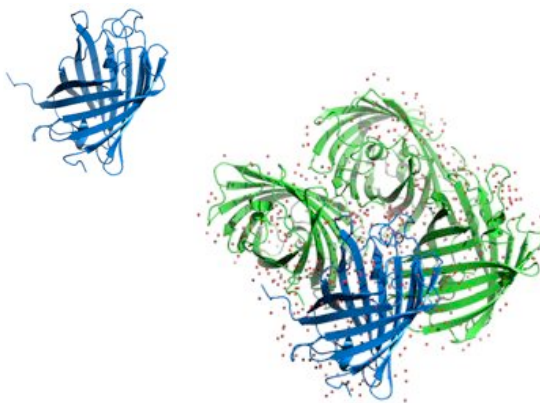


x-ray diffraction pattern (2.1 Å) for a SARS protease.

Source: Wikipedia @ [http://en.wikipedia.org/wiki/Image:X-ray\\_diffraction\\_pattern\\_3clpro.jpg](http://en.wikipedia.org/wiki/Image:X-ray_diffraction_pattern_3clpro.jpg)

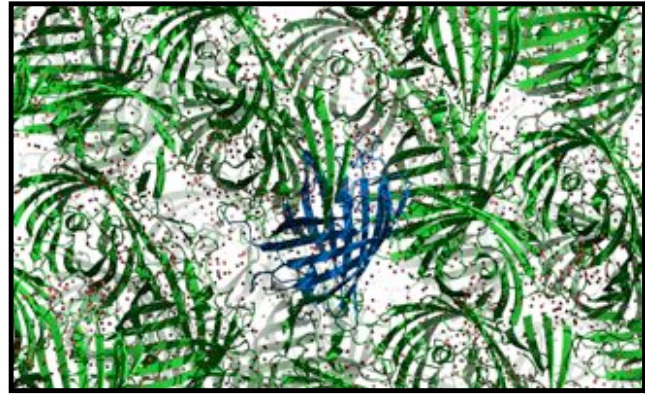
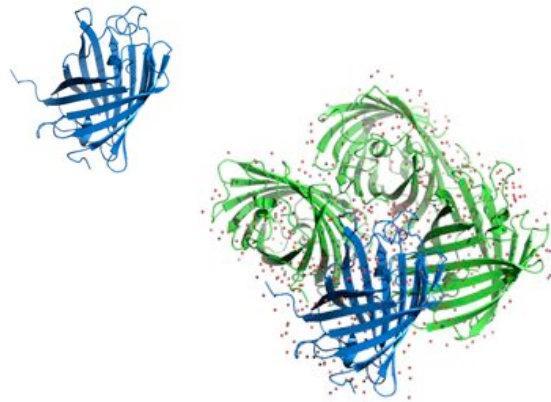
- The amplitude and phases of the diffraction data can be used to construct an electron density map of the unit cell (repeating unit of the crystal).
- The electron density map can contain errors with the quality of the map being dependent on the resolution of the diffraction data.
- Resolution reflects uniformity of organization of crystal and is reported in terms of Å.
  - Low resolution  $\geq 5\text{Å}$
  - Medium resolution  $\approx 3\text{Å}$
  - High resolution  $\leq 2\text{Å}$
- Constructing model involves:
  1. Determining how polypeptide chain weaves through density map.
  2. Fit side chains based on sequence to density.
- Usually results in multiple arrangements that are compatible with the density map as well as discontinuous regions.
- Generate best fit to map.

## Building the Model



# Building the Model

- The initial model can contain errors, which can be removed through crystallographic refinement.
- Involves comparing acquired diffraction amplitudes to theoretical values generated using a hypothetical crystal based on the model.
- Difference between the theoretical and acquired amplitudes expressed as an R factor.
  - 0.0 = exact agreement
  - ~0.59 = total disagreement
  - R factor usually 0.15-0.2 for a well determined structure.
- Residual disagreement usually due to errors in data and can reflect slight variations in conformation and/or lack of homogeneity in the crystal.
- B factor (temperature factor):
  - $\leq 20$  when structure is well defined
  - $\geq 40$  in local region may reflect conformational flexibility.
- High resolution structures ( $\sim 2\text{\AA}$ ) associated with low incidence of error.



# Nuclear Magnetic Resonance (NMR)

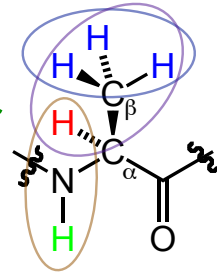
- Biologically relevant nuclei with magnetic moments (spin):  
 $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$  and  $^{31}\text{P}$
- In theory, it should be possible to obtain unique signals for each proton in a protein molecule.
- 1D spectra do not offer sufficient resolution. Differences in the chemical shifts of these protons are usually not within the resolving power of the experiment.
- Problem has been overcome\* by modern 2D NMR techniques/experiments.
- COSY (correlation spectroscopy): gives peaks for interactions between protons separated by one or two other atoms.
- NOESY (from NOE - nuclear overhouser effect): gives peaks for interactions between nuclei that are in close proximity of each other.
- More experiments have been designed to provide additional information/constraints.



800 MHz NMR at Pacific Northwest National Lab

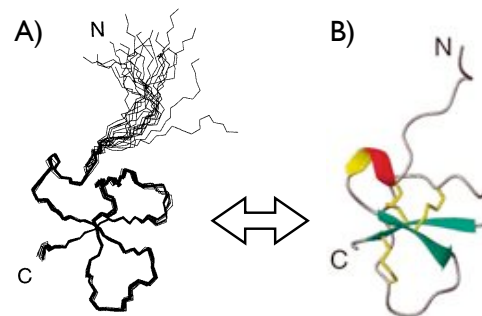
# Nuclear Magnetic Resonance (NMR)

- Sequential assignment:
  - Each amino acid has a specific set of covalently connected protons associated with it and results in a unique set of cross-peaks in COSY spectra.
  - COSY makes it possible to assign protons to specific amino acids.
  - NOESY spectra can be used to get sequence information.
  - Interactions between protons of sequentially adjacent residues (specifically main-chain N-H<sub>i+1</sub> and protons on the N, C<sub>α</sub> and C<sub>β</sub> positions of residue *i*).
  - Correlating COSY and NOESY data theoretically makes it possible to assign the sequence of the polypeptide. In practice really only feasible to make assignments of di- and tripeptide segments.
  - Identified segments are compared to known sequence of the peptide in order to fill in sequence and assign residues in the spectra.



- Unlike crystallography, NMR does not directly provide 3D structure. It instead provides distance constraints and identifies specific spatial relationships.
- Generates a list of distance constraints between specific protons of one amino acid and those of another amino acid in the protein. (long list)
- Distance constraints usually divided into three intervals within a range of 1.8-5 Å, depending the NOE peak intensity.
- Secondary structure can be identified based on the distance constraints ( $\alpha$  helices and  $\beta$  sheets have very specific sets of interactions <5Å).
- Models of the 3D structure of the protein can also be derived from these distance constraints.
- Usually results in a population of related possible structures rather than a single unique structure. Represent structures and structural permutations that are compatible with the set of distance constraints.

## NMR and Protein Structure



- Structure of peptide from platypus venom:
- Ensemble of 20 structures based on NMR constraints
  - "Cartoon" illustration based on ensemble of structures.

From: Torres, A.M., et al. *Biochem. J.* **1999** (341) 785-794

- Primary concerns when attempting to solve the three dimensional structure of a protein using NMR.
- Upper limit of ~25 kD.
- Requires concentrated solutions of protein (1-2 mM) without aggregation.
- pH should be <6 in order to minimize <sup>2</sup>H exchange of amide protons with bulk solvent.

# NMR and Limitations

- Primary concerns when attempting to solve the three dimensional structure of a protein using NMR.
  - Upper limit of ~25 kD.
  - Requires concentrated solutions of protein (1-2 mM) without aggregation.
  - pH should be <6 in order to minimize  $^2\text{H}$  exchange of amide protons with bulk solvent.
- Frequently requires NMR  $\geq$  500 MHz.

# Protein Engineering and Design

- Protein Engineering: aims to alter the function and/or physical properties of an existing protein by mutating the gene for that protein.
- Protein Design: strives to design, *de novo*, a protein to fulfill a desired property or function.
- Distinction between them can be somewhat unclear at times.