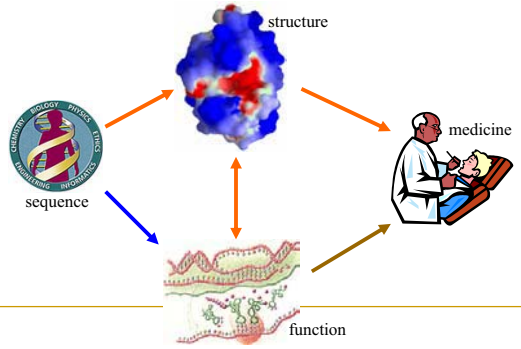
Protein structures

The basics of protein

- Proteins are linear polypeptide chains (one or more)
- Building blocks: 20 types of amino acids.
- Range from a few 10s-1000s
- They “fold” into varying three-dimensional shapes

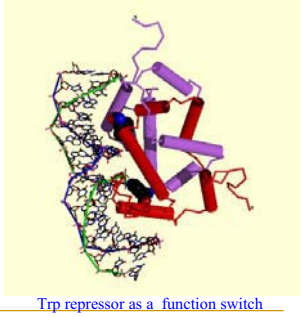
Relevance of Protein Structure in the Post-Genome Era



Structure-Function Relationship

Certain level of function can be found without structure. But a structure is a key to understand the detailed mechanism.

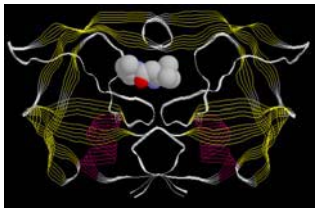
A predicted structure is a powerful tool for function inference.



Trp repressor as a function switch

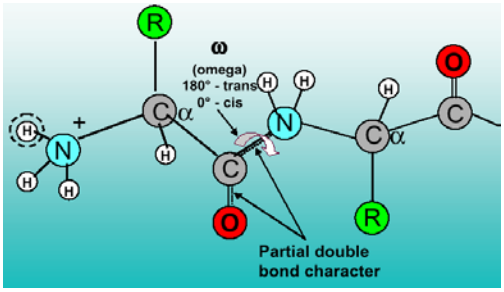
Structure-Based Drug Design

Structure-based rational drug design is a major method for drug discovery.



HIV protease inhibitor

Formation of polypeptide chain



Levels of protein structure

- Primary structure
- Secondary structure
- Tertiary structure
- Quaternary structure

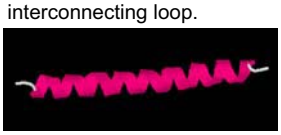
Primary structure

- This is simply the amino acid sequences of polypeptides chains (proteins).

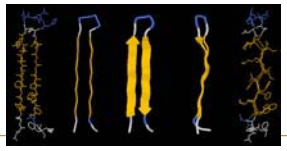
MHGAYRTPRSKTDAYGCQILETRAS

Secondary structure

- Local organization of protein backbone: α-helix, β-strand (groups of β-strands assemble into β-sheet), turn and interconnecting loop.



an α-helix



various representations and orientations of a two stranded β-sheet.

Secondary structure

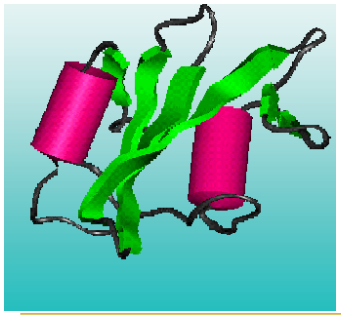
- Secondary structure is also a linear information and represented as string similar to amino acid sequence of proteins

AA sequence:	SGASDQSVSETYYIAAQCAPVGGQDA
Secondary structure:	HHHHHCCCEEEEEECCCEEEECCHHH

Tertiary structure

- Three-dimensional coordinates of the atoms of a chain is its *tertiary structure* (the structure of a single chain of a protein)
- *Quaternary structure*: describes how different changes are positioned relatively (the overall protein structure)

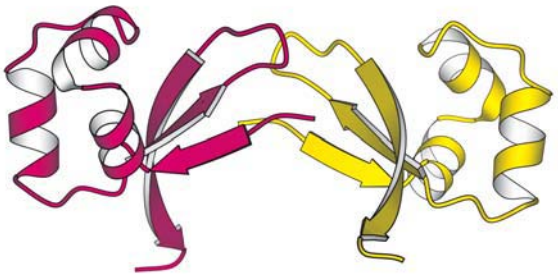
Tertiary structure

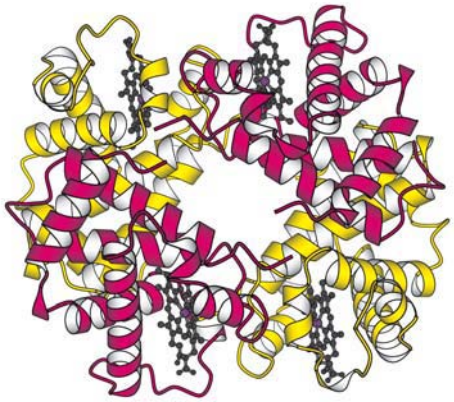


Packing the secondary structure elements into a compact spatial unit

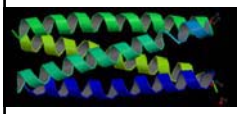
Quaternary structure

- Assembly of homo or heteromeric protein chains.
- Usually the functional unit of a protein, especially for enzymes

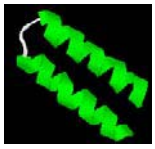




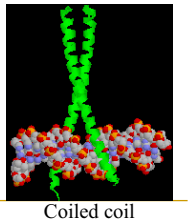
Structural Motifs



Four helix bundle



Helix-loop-helix



Coiled coil

- Primary and secondary structure are ONE-dimensional; Tertiary and quaternary structure are THREE-dimensional.
- "structure" usually refers to 3-D structure of protein.

PDB Files: the "header"

```
HEADER    OXIDOREDUCTASE(SUPEROXIDE ACCEPTOR)      13-JUL-94
COMPND    MANGANESE SUPEROXIDE DISMUTASE (E.C.1.15.1.1) COMPLEXED
COMPND    2 WITH AZIDE
SOURCE    (THERMUS THERMOPHILUS, HB8)
AUTHOR    M.S.LAH,M.DIXON,K.A.PATTRIDGE,W.C.STALLINGS,J.A.FEE,
AUTHOR    2 M.L.LUDWIG
REVDAT    2   15-MAY-95
REVDAT    1   15-OCT-94
JRNL      AUTH   M.S.LAH,M.DIXON,K.A.PATTRIDGE,W.C.STALLINGS,
JRNL      AUTH 2 J.A.FEE,M.L.LUDWIG
JRNL      TITL   STRUCTURE-FUNCTION IN E. COLI IRON SUPEROXIDE
JRNL      TITL 2 DISMUTASE: COMPARISONS WITH THE MANGANESE ENZYME
JRNL      TITL 3 FROM T. THERMOPHILUS
JRNL      REF    TO BE PUBLISHED
REMARK    1  AUTH   M.L.LUDWIG,A.L.METZGER,K.A.PATTRIDGE,W.C.STALLINGS
REMARK    1  TITL   MANGANESE SUPEROXIDE DISMUTASE FROM THERMUS
REMARK    1  TITL 2 THERMOPHILUS.  A STRUCTURAL MODEL REFINED AT 1.8
REMARK    1  TITL 3 ANGSTROMS RESOLUTION
REMARK    1  REF    J.MOL.BIOL.                   V. 219   335 1991
REMARK    1  REFN   ASTM JMOBAK  UK ISSN 0022-2836
REMARK    1 REFERENCE 2
REMARK    1  AUTH   W.C.STALLINGS,C.BULL,J.A.FEE,M.S.LAH,M.L.LUDWIG
REMARK    1  TITL   IRON AND MANGANESE SUPEROXIDE DISMUTASES:
REMARK    1  TITL 2 CATALYTIC INFERENCES FROM THE STRUCTURES
```

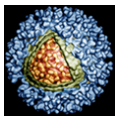
PDB Files: the coordinates

Atom & Residue | **XYZ Coordinates**

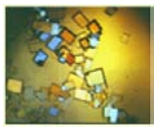
```
ATOM      1  N   PRO A   1      10.846  26.225 -13.938  1.00 30.15      1MNG 192
ATOM      2  CA  PRO A   1      12.063  25.940 -14.715  1.00 28.55      1MNG 193
ATOM      3  C   PRO A   1      12.061  26.809 -15.946  1.00 26.55      1MNG 194
ATOM      4  O   PRO A   1      11.151  27.612 -16.176  1.00 26.17      1MNG 195
ATOM      5  CB  PRO A   1      12.010  24.474 -15.162  1.00 30.21      1MNG 196
ATOM      6  CG  PRO A   1      11.044  23.902 -14.231  1.00 31.38      1MNG 197
ATOM      7  CD  PRO A   1       9.997  25.028 -14.008  1.00 31.86      1MNG 198
ATOM      8  N   TYR A   2      13.050  26.576 -16.777  1.00 23.36      1MNG 199
ATOM      9  CA  TYR A   2      13.197  27.328 -17.983  1.00 22.11      1MNG 200
ATOM     10  C   TYR A   2      12.083  27.050 -19.032  1.00 21.02      1MNG 201
ATOM     11  O   TYR A   2      11.733  25.895 -19.264  1.00 21.68      1MNG 202
ATOM     12  CB  TYR A   2      14.579  26.999 -18.523  1.00 20.16      1MNG 203
ATOM     13  CG  TYR A   2      14.905  27.662 -19.832  1.00 19.42      1MNG 204
ATOM     14  CD1 TYR A   2      14.516  27.092 -21.038  1.00 18.28      1MNG 205
ATOM     15  CD2 TYR A   2      15.610  28.864 -19.875  1.00 19.69      1MNG 206
ATOM     16  CE1 TYR A   2      14.813  27.696 -22.233  1.00 19.13      1MNG 207
ATOM     17  CE2 TYR A   2      15.924  29.465 -21.070  1.00 19.25      1MNG 208
ATOM     18  CZ  TYR A   2      15.515  28.863 -22.251  1.00 19.25      1MNG 209
ATOM     19  OH  TYR A   2      15.857  29.417 -23.448  1.00 21.67      1MNG 210
ATOM     20  N   PRO A   3      11.583  28.094 -19.731  1.00 19.90      1MNG 211
ATOM     21  CA  PRO A   3      11.912  29.520 -19.665  1.00 18.36      1MNG 212
```

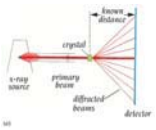
Experimental techniques for structure determination

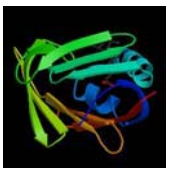
- X-ray Crystallography
- Nuclear Magnetic Resonance spectroscopy (NMR)
- Electron Microscopy/Diffraction

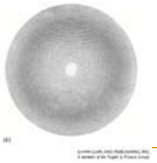


X-ray Crystallography



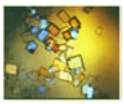






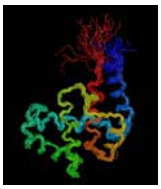
X-ray Crystallography..

- From small molecules to viruses
- Information about the positions of individual atoms
- Limited information about dynamics
- Requires crystals



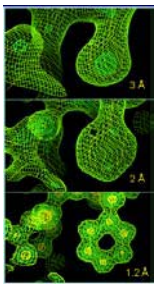
NMR

- Limited to molecules up to ~50kDa (good quality up to 30 kDa)
- Information about distances between pairs of atoms
  - A 2-d resonance spectrum with off-diagonal peaks
- Requires soluble, non-aggregating material



What does resolution mean for the structure?

- Low resolution, 6Å
- Medium resolution, 3Å
- High resolution, 1.5Å



Secondary structure prediction

- Given a protein sequence (primary structure)

GHWIATRGQLIREAYEDYRHFSSECPFIP

- Predict its secondary structure content (C=coils H=Alpha Helix E=Beta Strands)

CEEEEECHHHHHHHHHHHCCCHHCCCCCC

Why Secondary Structure Prediction?

- Easier problem than 3D structure prediction (more than 40 years of history).
- Accurate secondary structure prediction can be an important information for the tertiary structure prediction
- Improving sequence alignment accuracy
- Protein function prediction/classification

Prediction Methods

- Statistical methods
  - Chou-Fasman method, GOR I-IV
- Nearest neighbors
  - NNSSP, SSPAL
- Neural network
  - PHD, Psi-Pred, J-Pred
- Support vector machines
- Hidden Markov Models

Chou-Fasman method

- Compute parameters for amino acids
  - Preference to be in
    - alpha helix: P(a)
    - beta sheet: P(b)
    - Turn: P(turn)
  - Frequencies with which the amino acid is in the 1st, 2nd, 3rd, and 4th position of a turn: f(i), f(i+1), f(i+2), f(i+3).
- Use a sliding window

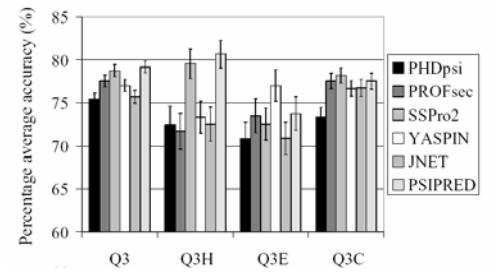
SSE prediction by Chou-Fasman

- Alpha-helix prediction
  - Find all regions where 4 of the 6 amino acids in window have P(a) > 100.
  - Extend the region in both directions unless 4 consecutive residues have P(a) < 100.
  - If Σ P(a) > Σ P(b) then the region is predicted to be alpha-helix.
- Beta-sheet prediction is analogous.
- Turn prediction
  - Compute P(t) = f(i) * f(i+1) * f(i+2) * f(i+3) for 4 consecutive residues.
  - Predict a turn if
    - P(t) > 0.000075 (check)
    - The average P(turn) > 100
    - Σ P(turn) > Σ P(a) and Σ P(turn) > Σ P(b)

GOR method

- Use a sliding window of 17 residues
- Compute the frequencies with which each amino acid occupies the 17 positions in helix, sheet, and turn.
- Use this to predict the SSE probability of each residue.

Performance of SSE prediction



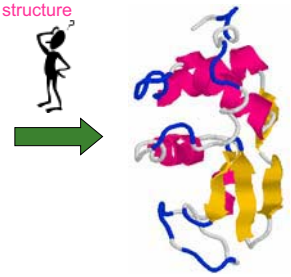
A Simple and Fast Secondary Structure Prediction Method using Hidden Neural Networks
Kuang Lin, Victor A. Simossis, William R. Taylor, Jaap Heringa, Bioinformatics Advance Access published September 17, 2004

Protein Folding Problem

A protein folds into a unique 3D structure under the physiological condition: determine this structure

Lysozyme sequence:

```
KVFGRCELAA AMKRHGLDNY
RGYSLGNWVC AAKFESNFNT
QATNRNTDGS TDYGILQINS
RWWCNDGRTP GSRNLCNIPC
SALLSSDITA SVNCAKKIVS
DGNGMNAWVA WRNRCKGTDV
QAWIRGCRL
```



Levinthal's paradox

- Consider a 100 residue protein. If each residue can have only 3 conformations, there are $3^{100} = 5 \times 10^{47}$ possible conformations.
  - If it takes 10^{-13}s to convert from 1 structure to another, exhaustive search would take 1.6×10^{27} years!
- Folding must proceed by progressive stabilization of intermediates.

Forces driving protein folding

- It is believed that *hydrophobic collapse* is a key driving force for protein folding
  - Hydrophobic core
  - Polar surface interacting with solvent
- Minimum volume (no cavities)
- Disulfide bond formation stabilizes structure
- Hydrogen bonds
- Polar and electrostatic interactions

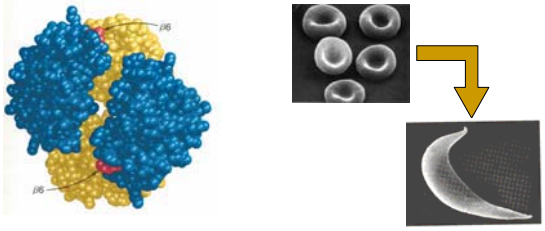
Folding help

- Proteins are, in fact, only marginally stable
  - Native state is little more stable than the unfolded form
- Many proteins help in folding
  - Protein disulfide isomerase – catalyzes shuffling of disulfide bonds
  - Chaperones –alter protein structures and (in theory) unfold misfolded proteins

Effect of a single mutation

- Hemoglobin is the protein in red blood cells (erythrocytes) responsible for binding oxygen.
- The mutation E→V in the β chain replaces a charged Glu by a hydrophobic Val on the surface of hemoglobin
- The resulting "sticky patch" causes hemoglobin to stick together and form fibers which deform the red blood cell and do not carry oxygen efficiently
- Sickle cell anemia was the first identified molecular disease

Sickle Cell Anemia



Sequestering hydrophobic residues in the protein core protects proteins from hydrophobic agglutination (sticking together).

Protein Structure Prediction

- *Ab-initio* techniques
- Homology modeling
  - Sequence-sequence comparison
- Protein threading
  - Sequence-structure comparison