# Three Lectures

- Structure Determination
- Energetics
- Structure Prediction


# Protein Structure Determination

**Dr. Fayyaz ul Amir Afsar Minhas**
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

# Protein Structure

- How is the structure of a protein determined?
  - Sequencing
  - Tertiary Structure Determination
    - X-ray crystallography
    - NMR Spectroscopy
    - Cryo-Electron microscopy
- Structural Alignment
- Structure Prediction
  - Secondary and Super-Secondary Structure and Properties Prediction
  - Tertiary Structure Prediction
  - Structural Dynamics
- Protein Interactions
  - Interaction Prediction
  - Binding site prediction
    - Docking

# Protein Sequencing

- Protein Sequencer
  - They work by tagging and removing one amino acid at a time, which is analysed and identified. This is done repetitively for the whole polypeptide, until the whole sequence is established.
  - This method has generally been replaced by nucleic acid technology, and it is often easier to identify the sequence of a protein by looking at the DNA that codes for it.
- Mass-Spectrometry
- Edman Degradation

A Beckman-Coulter Porton LF3000G protein sequencing machine

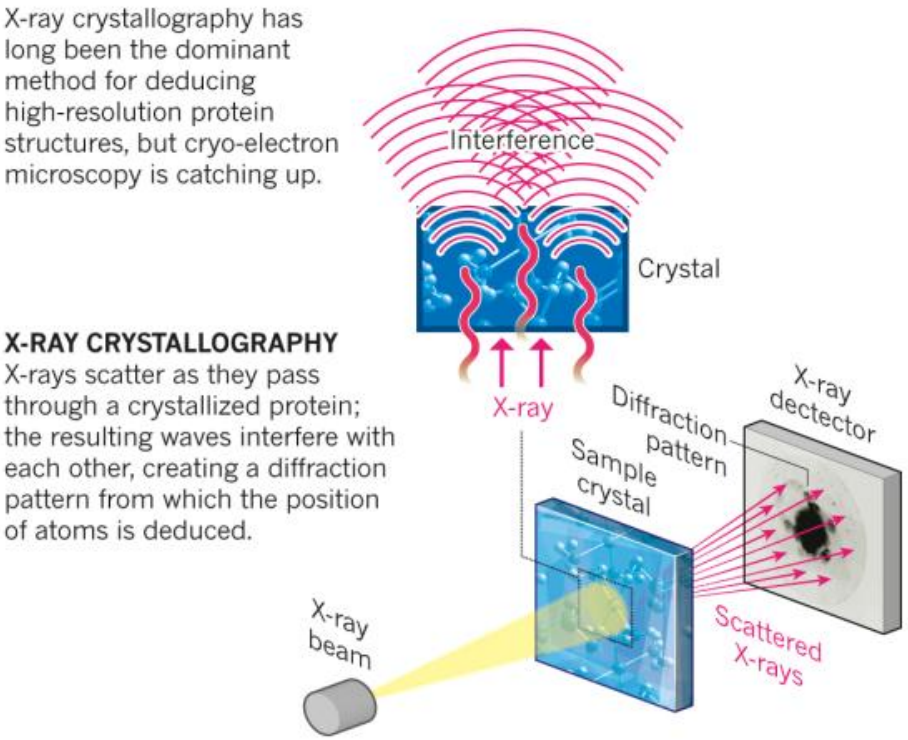
https://en.wikipedia.org/wiki/Protein_sequencing

# X-ray Crystallography

- 1960 John Kendrew determined the first structure of a protein using X-ray diffraction
- 2014 has been declared by the UN as the international year of crystallography
  - 1914 Max von Laue won the Nobel prize for discovering the diffraction of X-rays by crystals
- Most accurate method today for structural determination of proteins

# Steps

- Isolation and purification
  - May have to remove flexible regions
- Crystallization
  - Start from highly concentrated solution of the protein
  - Weeks to Months required to get crystals of sufficient size (20-300 micrometers)



# X-ray Crystallography

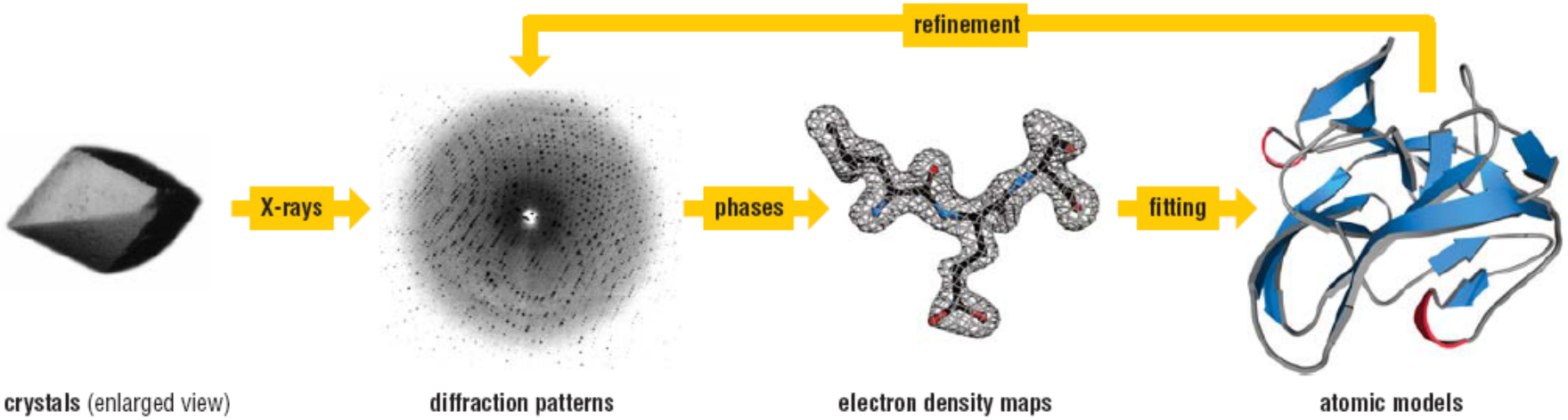


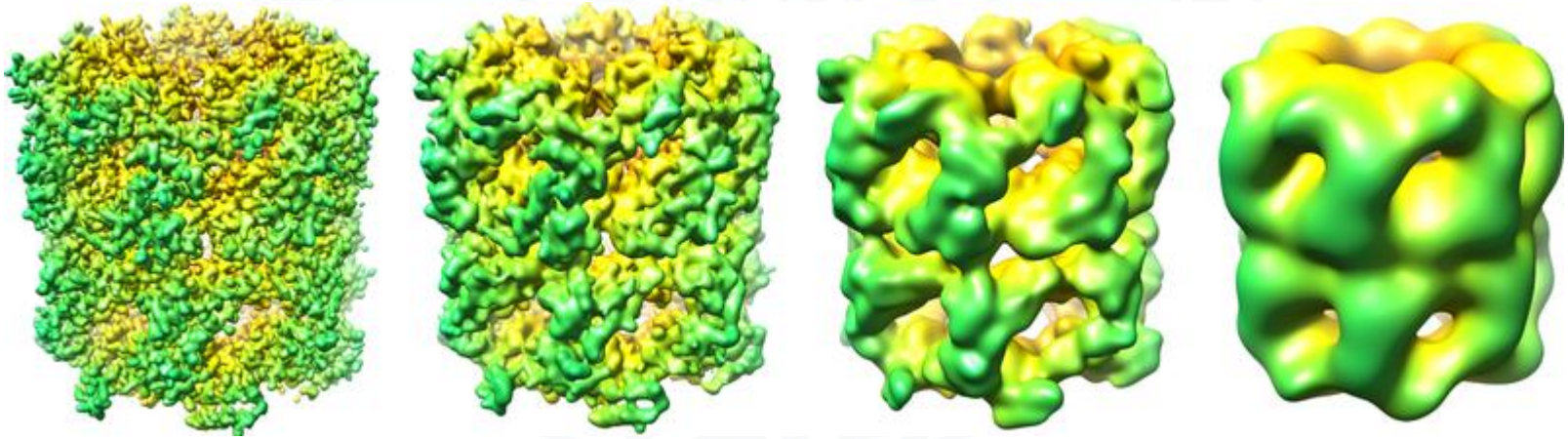

**Figure 5-3 Structure determination by X-ray crystallography** The first step in structure determination by X-ray crystallography is the crystallization of the protein. The source of the X-rays is often a synchrotron and in this case the typical size for a crystal for data collection may be $0.3 \times 0.3 \times 0.1$ mm. The crystals are bombarded with X-rays which are scattered from the planes of the crystal lattice and are captured as a diffraction pattern on a detector such as film or an electronic device. From this pattern, and with the use of reference—or phase—information from labeled atoms in the crystal, electron density maps (shown here with the corresponding peptide superimposed) are computed for different parts of the crystal. A model of the protein is constructed from the electron density maps and the diffraction pattern for the modeled protein is calculated and compared with the actual diffraction pattern. The model is then adjusted—or refined—to reduce the difference between its calculated diffraction pattern and the pattern obtained from the crystal, until the correspondence between model and reality is as good as possible. The quality of the structure determination is measured as the percentage difference between the calculated and the actual pattern.

# Information obtained from crystallography

- Molecular structure
- Deviation
  - Atoms have kinetic energy (i.e., they move)
  - The location obtained by crystallography is thus the average location
  - We get the B-factor (aka temperature factor)
    - Measure of deviation of an atom from its average location
      - Can be caused by thermal motions
      - Also by conformational changes
    - B < 45 indicate ordered atoms
  - See in pyMOL

# Information obtained from crystallography

- Resolution
  - Smallest separation between atoms diffracting the X-rays



4Å 8Å 16Å 32Å

http://commons.wikimedia.org/wiki/File:GroEL_resolution_series.png

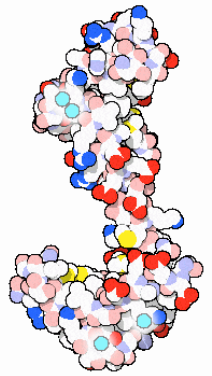
# Problems

- Preparation of the protein
  - X-rays can damage the sample so steps needed to be taken to prevent that
  - Crystallization can be very difficult
  - May have to change the protein significantly for it to be able to find crystals
    - The structure in the crystal maybe different from its real structure
- Effects of crystallization
  - Structure determined in unnatural conditions
- Hydrogens cannot be “seen” due to their low electron density (Neutron scattering)
- Crystal packing interfaces in macromolecular assemblies

# Calmodulin

- 1CLL: Calcium bound calmodulin (holo)

**August 2003: Calmodulin**



Calcium is the most plentiful mineral element found in your body, with phosphorous coming in second. This probably doesn't come as a surprise, since your bones are strengthened and supported by about two kilograms of calcium and phosphorous. Your body also uses a small amount of calcium, in the form of calcium ions, to perform more active duties. Calcium ions play essential roles in cell signaling, helping to control processes such as muscle contraction, nerve signaling, fertilization and cell division. Through the action of calcium pumps and several kinds of calcium binding proteins, cells keep their internal calcium levels 1000–10,000 times lower than the calcium levels in the blood. Thus when calcium is released into cells, it can interact with calcium sensing proteins and trigger different biological effects, causing a muscle to contract, releasing insulin from the pancreas, or blocking the entry of additional sperm cells once an egg has been fertilized.

**Sensing Calcium**

As its name suggests, calmodulin is a CALcium MODULated protelN. It is abundant in the cytoplasm of all higher cells and has been highly conserved through evolution. Calmodulin acts as an intermediary protein that senses calcium levels and relays signals to various calcium–sensitive enzymes, ion channels and other proteins. Calmodulin is a small dumbbell–shaped protein composed of two globular domains connected together by a flexible linker. Each end binds to two calcium ions. PDB entry 3cln, shown here, has all four sites filled with calcium ions and the linker has formed a long alpha helix separating the two calcium–binding domains.

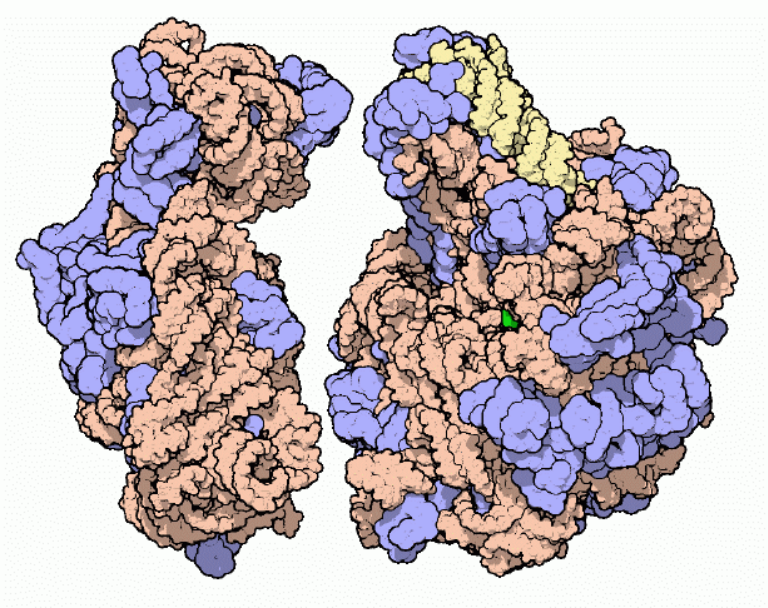
**Calmodulin Look–alikes**

Many different proteins are sensitive to calcium levels inside (and outside) cells. In the late 1960's, before the discovery of calmodulin, troponin C (see, for instance, PDB entry 1tcf) was the first protein shown to be sensitive to calcium. Troponin C senses rising calcium levels and triggers muscle contraction. The structures of troponin C and calmodulin are remarkably similar, the major difference being the length of the linker connecting the two calcium–binding globular domains. The calcium–binding region of the protein, shown in detail in a later section, is almost identical. This motif has since been found in dozens of other calcium–sensitive proteins.

http://www.rcsb.org/pdb/101/motm.do?momID=44

# Examples

- Ribosome



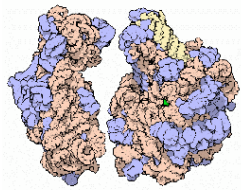
**October 2000: Ribosome**

**The Protein Factory**

Protein synthesis is the major task performed by living cells. For instance, roughly one third of the molecules in a typical bacterial cell are dedicated to this central task. Protein synthesis is a complex process involving many molecular machines. You can look at many of these molecules in the PDB, including DNA, DNA polymerases, and RNA polymerases; a host of repressors, DNA repair enzymes, topoisomerases, and histones; tRNA and acyl-tRNA synthetases; and molecular chaperones. This month, for the first time, you can also look at the factory of protein synthesis in atomic detail.

**An Elusive Structure**

The ribosome has been under the scrutiny of scientists for decades. Electron microscopy has yielded an increasingly detailed view over the years, defining the overall shape of individual ribosomes and differences in this shape for ribosomes from different species. More recently, detailed electron micrograph reconstructions have studied the interaction of ribosomes with messenger RNA, transfer RNA and the protein elongation factors. This legacy of morphological work lays the groundwork on which the atomic structures may be understood.



Ribosomes are composed of two subunits: a large subunit, shown on the right, and a small subunit, shown on the left. Of course, the term "small" is used in a relative sense here: both the large and the small subunits are huge compared to a typical protein. Both subunits are composed of long strands of RNA, shown here in orange and yellow, dotted with protein chains, shown in blue.

When synthesizing a new protein, the two subunits lock together with a messenger RNA trapped in the space between. The ribosome then walks down the messenger RNA three nucleotides at a time, building a new protein piece-by-piece.

http://www.rcsb.org/pdb/101/motm.do?momID=10

# NMR Spectroscopy

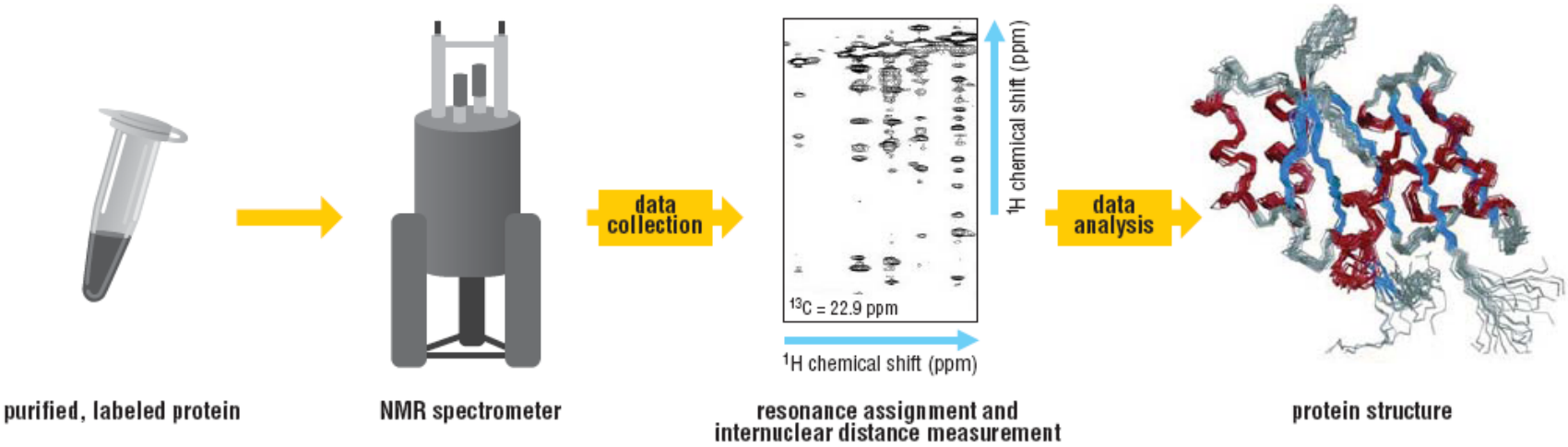



**Figure 5-4 Structure determination by NMR** For protein structure determination by NMR, a labeled protein is dissolved at very high concentration and placed in a magnetic field, which causes the spin of the hydrogen atoms to align along the field. Radio frequency pulses are then applied to the sample, perturbing the nuclei of the atoms which when they relax back to their original state emit radio frequency radiation whose properties are determined by the environment of the atom in the protein. This emitted radiation is recorded in the NMR spectrometer for pulses of differing types and durations (for simplicity, only one such record is shown here), and compared with a reference signal to give a measure known as the chemical shift. The relative positions of the atoms in the molecule are calculated from these data to give a series of models of the protein which can account for these data. The quality of the structure determination is measured as the difference between the different models.

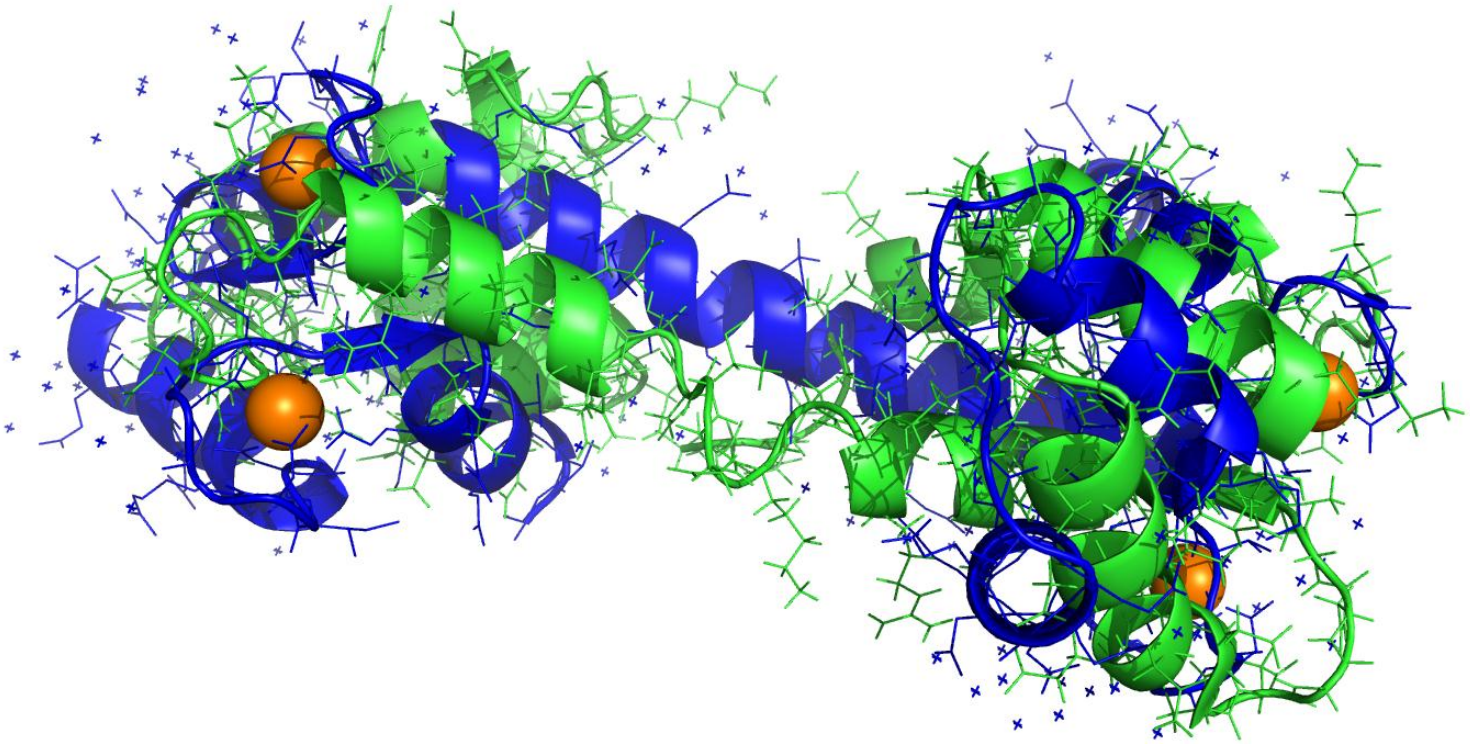
# NMR Spectroscopy

- No crystallization required
- Gives ensembles of structures (possible "states" of the molecule)
  - Protein dynamics can be studied at the nanosecond scale
- Detects Hydrogens

- Not possible for large proteins (>40kDa)

# NMR Structure of Calmodulin

- Structure of Calmodulin not bound to Calcium (apo-Calmodulin)



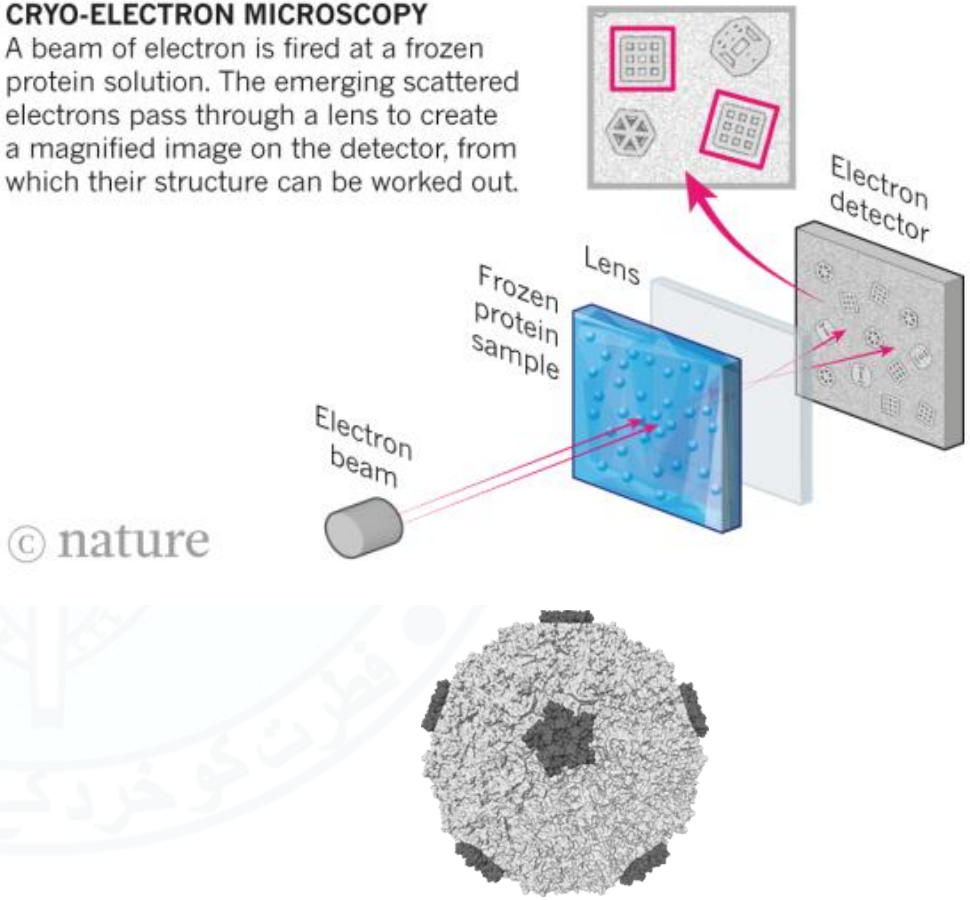
1CFD.pdb

# NMR Structure of Calmodulin

- Structure of Calmodulin not bound to Calcium (apo-Calmodulin): 30 States
  - PDB: 1DMO

# Cryo-Electron Microscopy

- Sample is studied at cryogenic temperatures
- No staining or fixing
- Lower resolution
  - 4.5 A
- In 2015, a 2.2 A resolution structure of an enzyme was determined




Callaway, Ewen. 2015. "The Revolution Will Not Be Crystallized: A New Method Sweeps through Structural Biology." *Nature* 525 (7568): 172–74. doi:10.1038/525172a.

# Calmodulin in complex with aquaporin

- 3J41

- Costs could slow the spread of the technology. Scheres estimates that the LMB spends around £3,000 per day running its cryo-EM facility, plus another £1,000 on electricity, most of it for computers needed to store and process the images. “You’re £4,000 per day lighter if you want to do this. That, for many places, is a very high cost”.

# Structure Related Computational Tasks

- Accessible surface area (ASA/RASA) estimation
- Protein energy calculations
- Protein Tertiary structure alignments
- Predictions
  - Protein secondary structure prediction
  - Protein contact map prediction
  - Protein torsion angle prediction
  - Protein ASA prediction
  - Protein disorder prediction
  - Protein tertiary structure prediction

# Structural Alignment

- Types
  - Global
  - Local
  - Multiple structure alignment
- RMSD
- Methods
  - DALI
  - FSSP
  - ProBiS
  - PyMOL

Structural alignment of thioredoxins from humans and the fly Drosophila melanogaster. The proteins are shown as ribbons, with the human protein in red, and the fly protein in yellow. Generated from PDB 3TRX and 1XWC.

http://en.wikipedia.org/wiki/Structural_alignment


# Protein Energetics

**Dr. Fayyaz ul Amir Afsar Minhas**
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

# Protein Energetics

- Why do proteins fold?
- What is the physics behind the behavior of proteins?
- How can we computationally model this physics?
- Meeting point of Physics, Chemistry and Biology

# Concept

- Probability of finding a (mole of a ) molecule in a particular state

$$p_A = w_A exp\left(-\frac{E_A}{RT}\right)$$

  - $E_A$: Total energy of the system (molecular kinetic, rotational, vibrational, interactions within and between molecules of the system)
  - $w_A$: Number of ways in which that total energy may be achieved or distributed, i.e., entropy
  - $T$: Temperature
  - $R$: Gas Constant

  - What states are more probable?
    - Based on energy
    - Based on entropy

https://en.wikipedia.org/wiki/Boltzmann_distribution

# Observing a molecule in a state

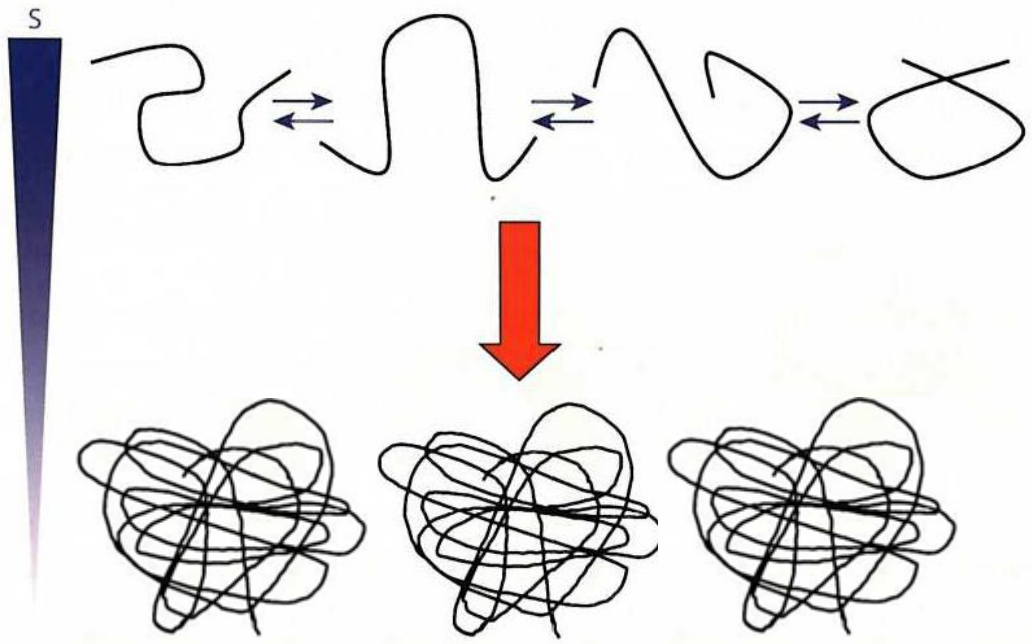
- Assume two states $A \rightleftharpoons B$
  - $p_A = w_A exp\left(-\frac{E_A}{RT}\right)$
  - $p_B = w_{\boldsymbol{B}} exp\left(-\frac{E_{\boldsymbol{B}}}{RT}\right)$
- The reaction constant is the ratio of the amount of B to that of A, i.e.,
  - $K = \frac{[\boldsymbol{B}]}{[\boldsymbol{A}]} = \frac{p_B}{p_A} = \frac{w_{\boldsymbol{B}} exp\left(-\frac{E_{\boldsymbol{B}}}{RT}\right)}{w_A exp\left(-\frac{E_A}{RT}\right)} = \frac{w_{\boldsymbol{B}}}{w_A} exp\left(-\frac{E_{\boldsymbol{B}} - E_A}{RT}\right)$
  - $\boldsymbol{\Delta G = \Delta H - T\,\Delta S}$
- For a biological system (which doesn't undergo changes in temperature or pressure)
  - $\boldsymbol{\Delta G = -RTln(K)}$
  - $\boldsymbol{\Delta E} = E_{\boldsymbol{B}} - E_A = \boldsymbol{\Delta H}$ = Change in internal energy
  - $\boldsymbol{\Delta S} = \boldsymbol{Rln}\left(\frac{w_{\boldsymbol{B}}}{w_A}\right)$ = Change in entropy

# Spontaneity of a reaction

- $\Delta G = \Delta H - T\,\Delta S$

- $\Delta G = -RT ln\left(\frac{p_B}{p_A}\right) > 0$ **implies** $p_B < p_A$
  - Reaction would not be spontaneous
- $\Delta G = -RT ln\left(\frac{p_B}{p_A}\right) < 0$ implies $p_B < p_A$
  - Reaction would be spontaneous

# Let's talk about proteins: entropy




Unfolded state
Ensemble of conformations existing in equilibrium which may be very different from each other

$$W_{unfolded}$$

Folded state
is a much smaller ensemble of conformations which exist in equilibrium. We can say that there is only one average conformation.

$$W_{folded}$$

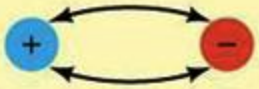
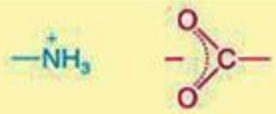
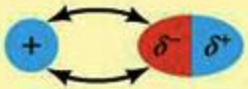
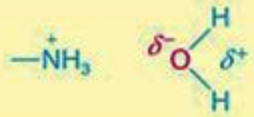
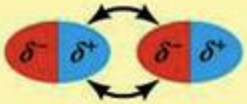
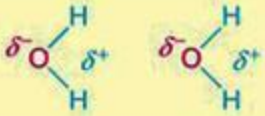
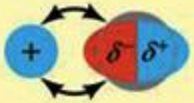
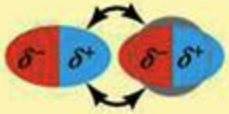
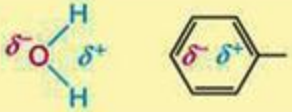
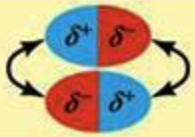
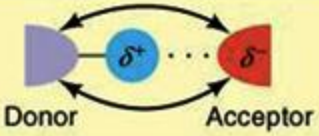
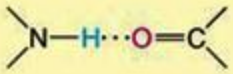
- Significant decrease in entropy
  - $\Delta S = Rln\left(\frac{W_{folded}}{W_{unfolded}}\right) < 0$

- Thus $\Delta S_{protein} < 0$
- For the process to be spontaneous
  - $\Delta G = \Delta H - T\Delta S < 0$
- Since $-T\Delta S > 0$, $\Delta H$ must be significantly negative
- Thus, we need to compute the change in energy of the system
  - For this, we need to calculate the energy of the system before and after folding

# Calculating change in Energy of a protein

- Binding Energies
  - Disulfide
  - Bound ions
  - Etc,
- Non-binding energies
  - Electrostatic
  - Hydrogen bonds
  - Van der Waals Forces

# Types of interactions

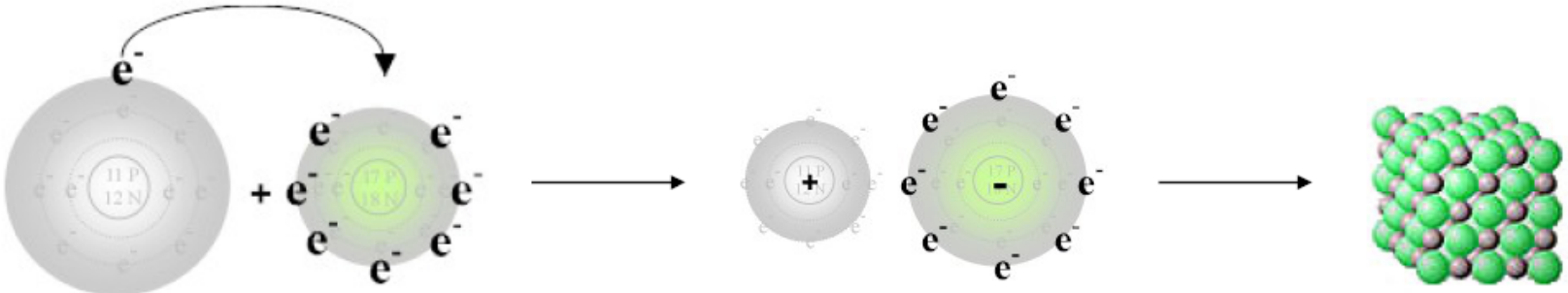
| Type of Interaction | Model | Example | Dependence of Energy on Distance |
|---|---|---|---|
| **(a) Charge–charge**<br>Longest-range force; nondirectional |  |  | $1/r$ |
| **(b) Charge–dipole**<br>Depends on orientation of dipole |  |  | $1/r^2$ |
| **(c) Dipole–dipole**<br>Depends on mutual orientation of dipoles |  |  | $1/r^3$ |
| **(d) Charge–induced dipole**<br>Depends on polarizability of molecule in which dipole is induced |  |  | $1/r^4$ |
| **(e) Dipole–induced dipole**<br>Depends on polarizability of molecule in which dipole is induced |  |  | $1/r^5$ |
| **(f) Dispersion (van der Waals)**<br>Involves mutual synchronization of fluctuating charges |  |  | $1/r^6$ |
| **(g) Hydrogen bond**<br>Charge attraction + partial covalent bond |  |  | Length of bond fixed |

# Electrostatic Interactions

Charged groups attract or repel each other. The force F of such an electrostatic interaction is given by Coulomb's law:

$$F = \frac{q_1 q_2}{D r^2}$$

$q_1$ and $q_2$ are the charges
$r$ is the distance
$D$ is the dielectric constant

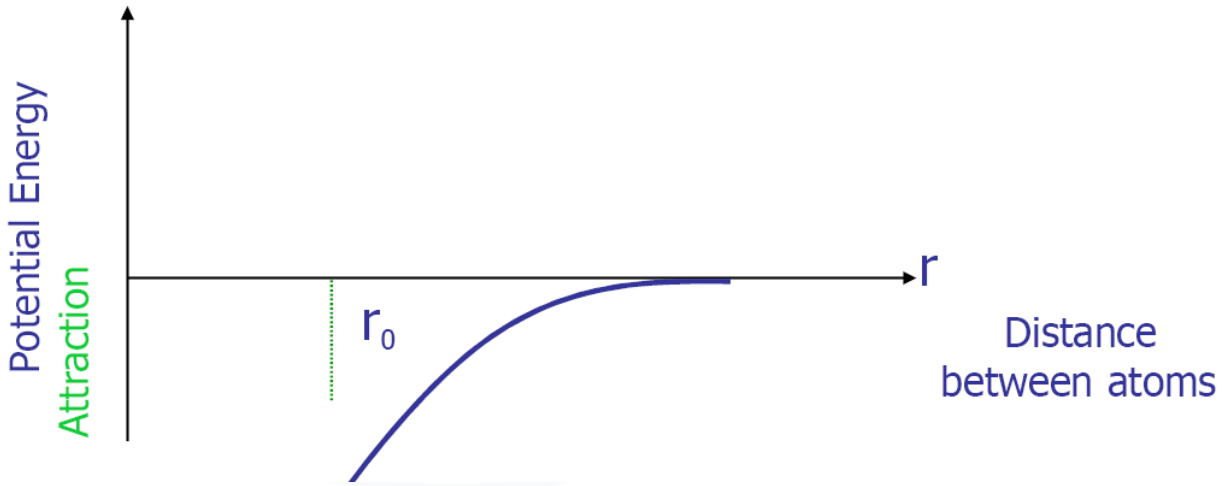



Coulomb's law is also used to determine interactions between uncharged, but polar atoms.

# Van der Waals Interactions

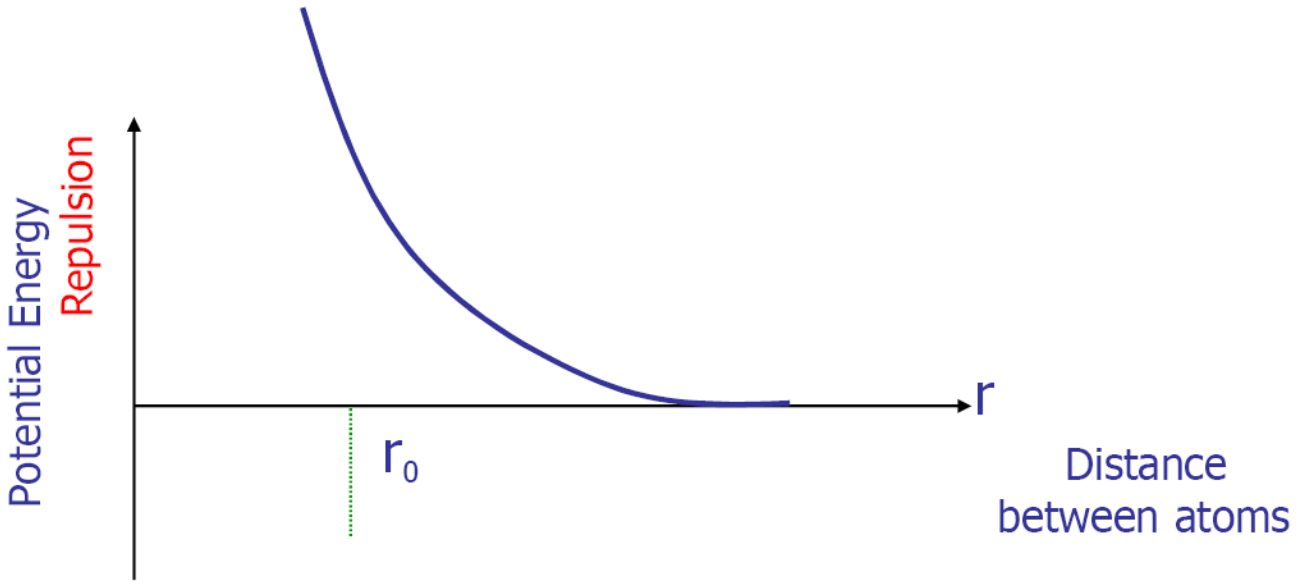
The distribution of electronic charges around an atom changes with time, and a transient asymmetry in the charges around one atom induces a similar asymmetry in the electron distribution around its neighboring atoms.

This is essentially an electrostatic interaction and results in a small distant-dependent ($R^{-6}$) attractive force.



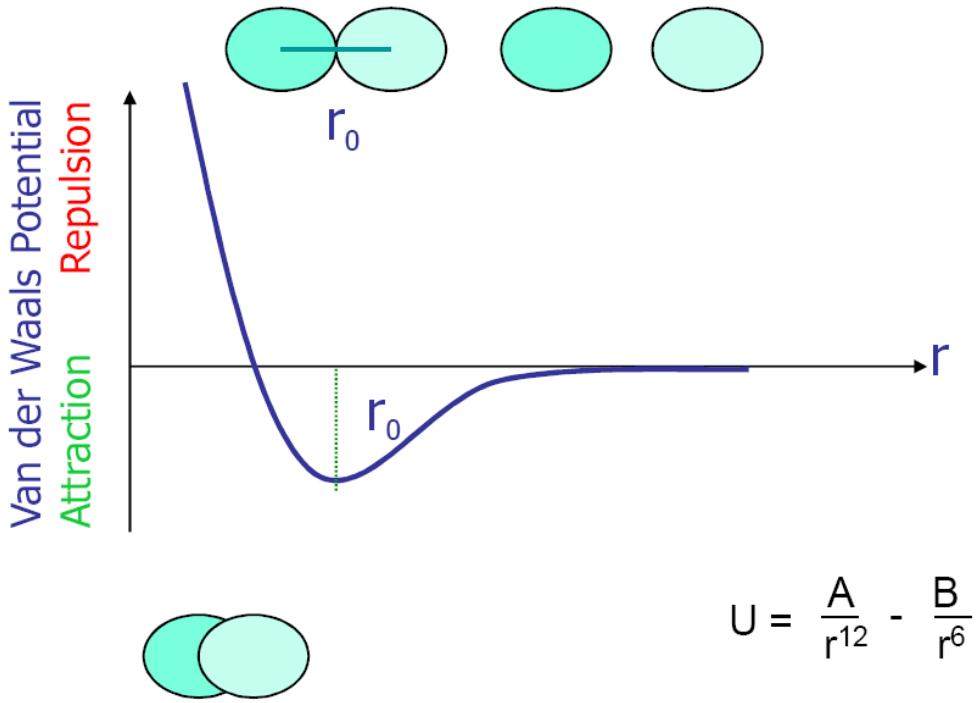
# Van der Waals Interactions

As atoms get too close, their electron clouds will clash, resulting a distant-dependent ($R^{-12}$) repulsive potential energy.



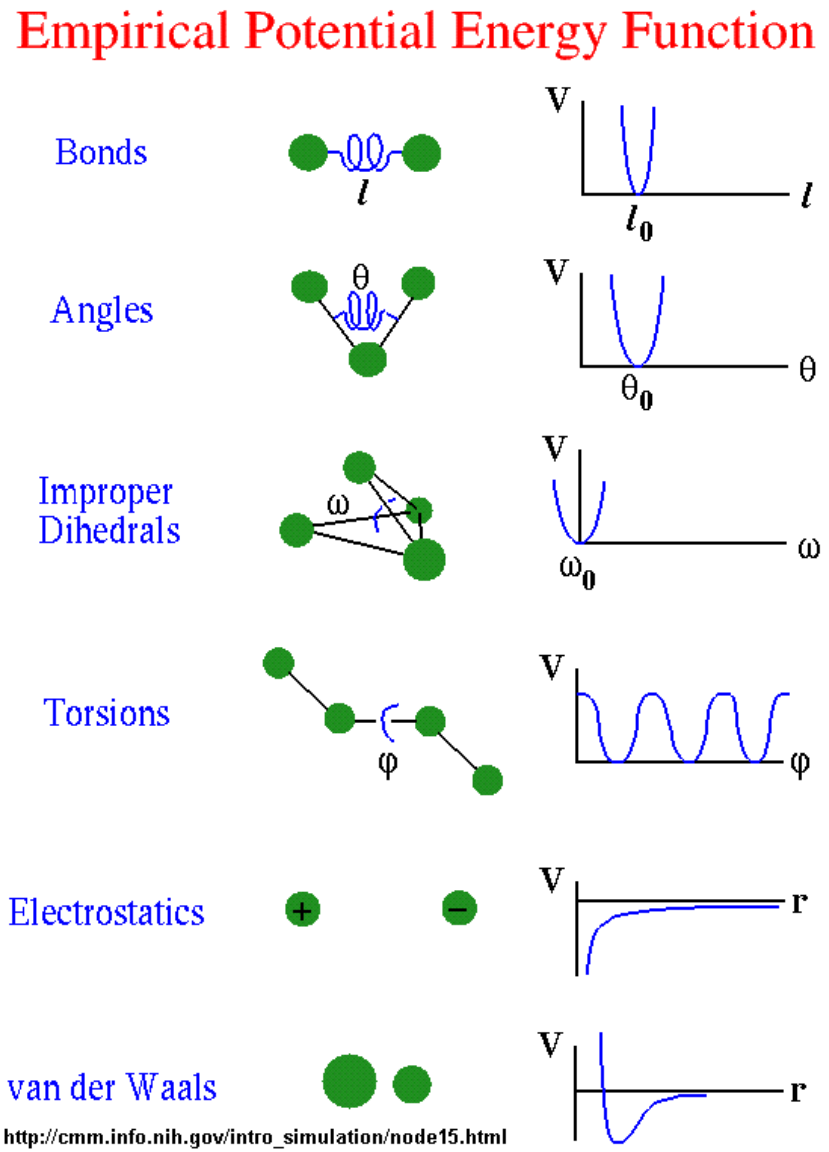
# Lennard-Jones Potential

- The attractive and repulsive terms can be summed together to describe a distance-dependent interatomic potential energy.



# Energy Calculations

- We can measure the potential energy of a molecule in molecular mechanics
  - $E_{total} = E_{bond} + E_{angle} + E_{torsion} + E_{electro} + E_{vdw}$
  - $\Sigma_{(i,j)\ E\ Sbond}\ k_{ij}^{b}(\alpha_{ijk}-\alpha_{ijk}^{0})^2$
    - Bond length
    - Bond Angle
    - Bond Torsion
  - van der Waals
    - $E_{vdw} = \Sigma_{(i,j)\ E\ Svdw}\ \varepsilon_{ij}\ [\ (\sigma_{ij}/r_{ij})^{12} - 2(\sigma_{ij}/r_{ij})^{6}]$
  - Electrostatics
    - $E_{electro} = \Sigma_{(i,j)\ E\ Selectro}\ (\ q_i q_j)/(e_{ij} r_{ij})$



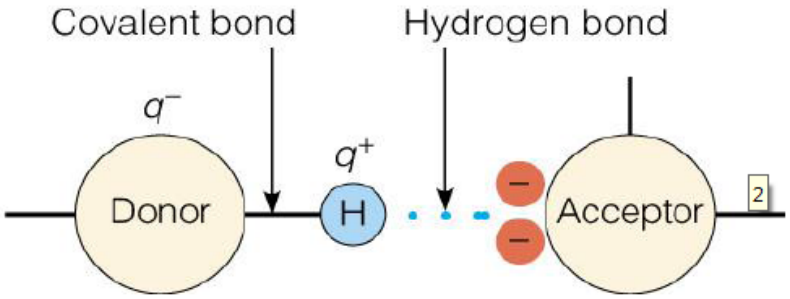

http://en.wikipedia.org/wiki/Potential_energy_of_protein

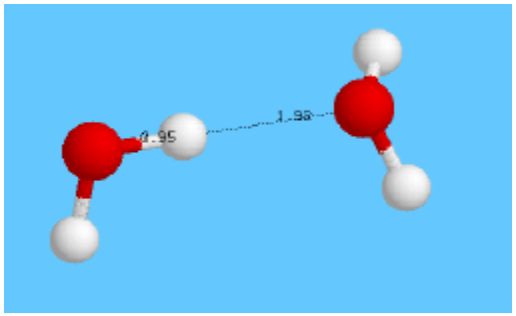
# Discrepencies

- A number of people have experimentally calculated the change in energy for some proteins due to folding
- However, it was found that this change in energy is small and is, by itself, insufficient to cause
- There must be some other factor
  - Entropy
  - But that decreases
- So what is the factor?
  - Water

To understand the structure of individual water molecules in ice and liquid water, you first need to understand the hydrogen bond, one of the major types of noncovalent interactions.

H-bond definition: interaction between a covalently bonded electropositive hydrogen atom on a donor group and a lone pair of non-bonded electrons on an acceptor group



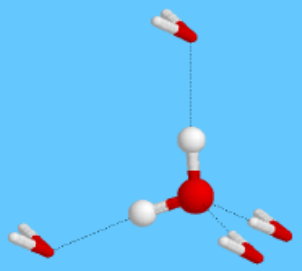

H-Bonding in water

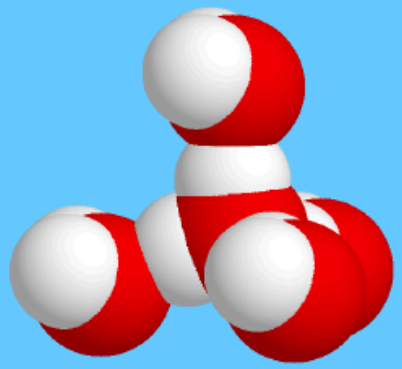
**Water is both an H-bond donor and acceptor!!**

The hydrogen bond has some features of covalent bonding: it is directional, strong, produces interatomic distances shorter than sum of van der Waals radii

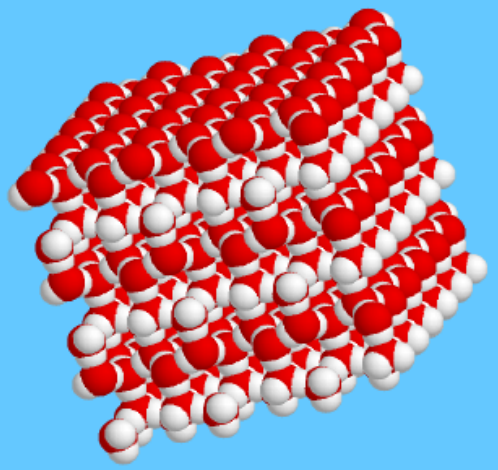
# Water forms H-bonds which are near perfect in Ice

Water has tetrahedral geometry in ice

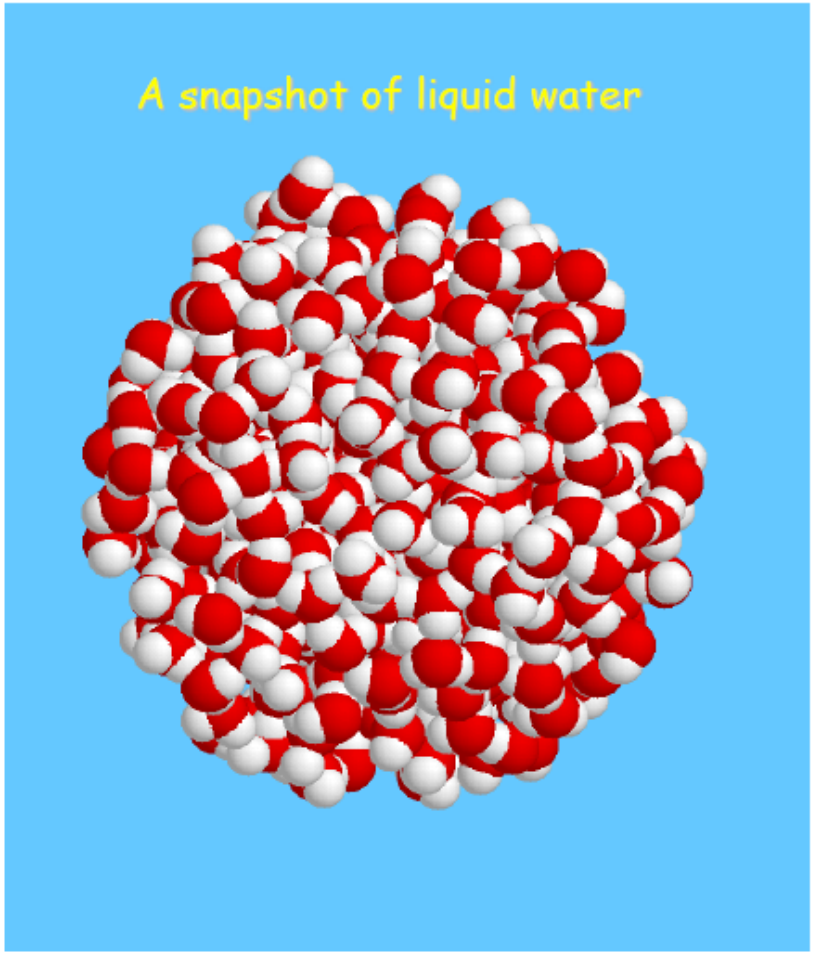




432 water molecules in ice



- Water is >99% H-bonded in ice
- When water freezes at 0°C, its volume increases by about 9% (and its density decreases). The packing density of ice at 0°C is 0.34. This means that 66% of the volume of ice is unoccupied because the packing is restricted by the hydrogen-bonding geometrical constraints.




Same 432 water molecules as liquid bulk water

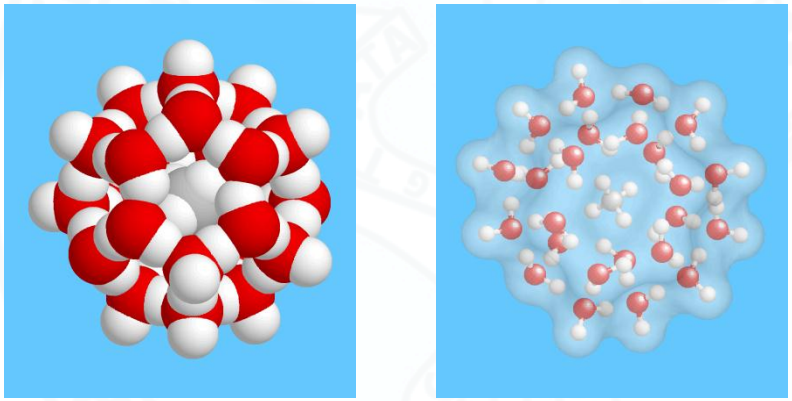
95% of the possible H-bonds are engaged in liquid water (and 85% in boiling water!)

In liquid water, hydrogen bonds are constantly stretching, bending, and breaking as the molecules rotate and jump around. **There is no regular stable tetrahedral geometry.** The average lifetime of a hydrogen bond is about one picosecond ($10^{-12}$ sec) in liquid water at 25°C.

It is this network of "flickering" hydrogen bonds that gives liquid water its unique properties. This flickering also accounts for the fact that while water is more dense than ice **(packing density is 0.37 at 4°C)**, the 'collapsed' structure is still 'open' because of the highly directional character of the H-bonds.

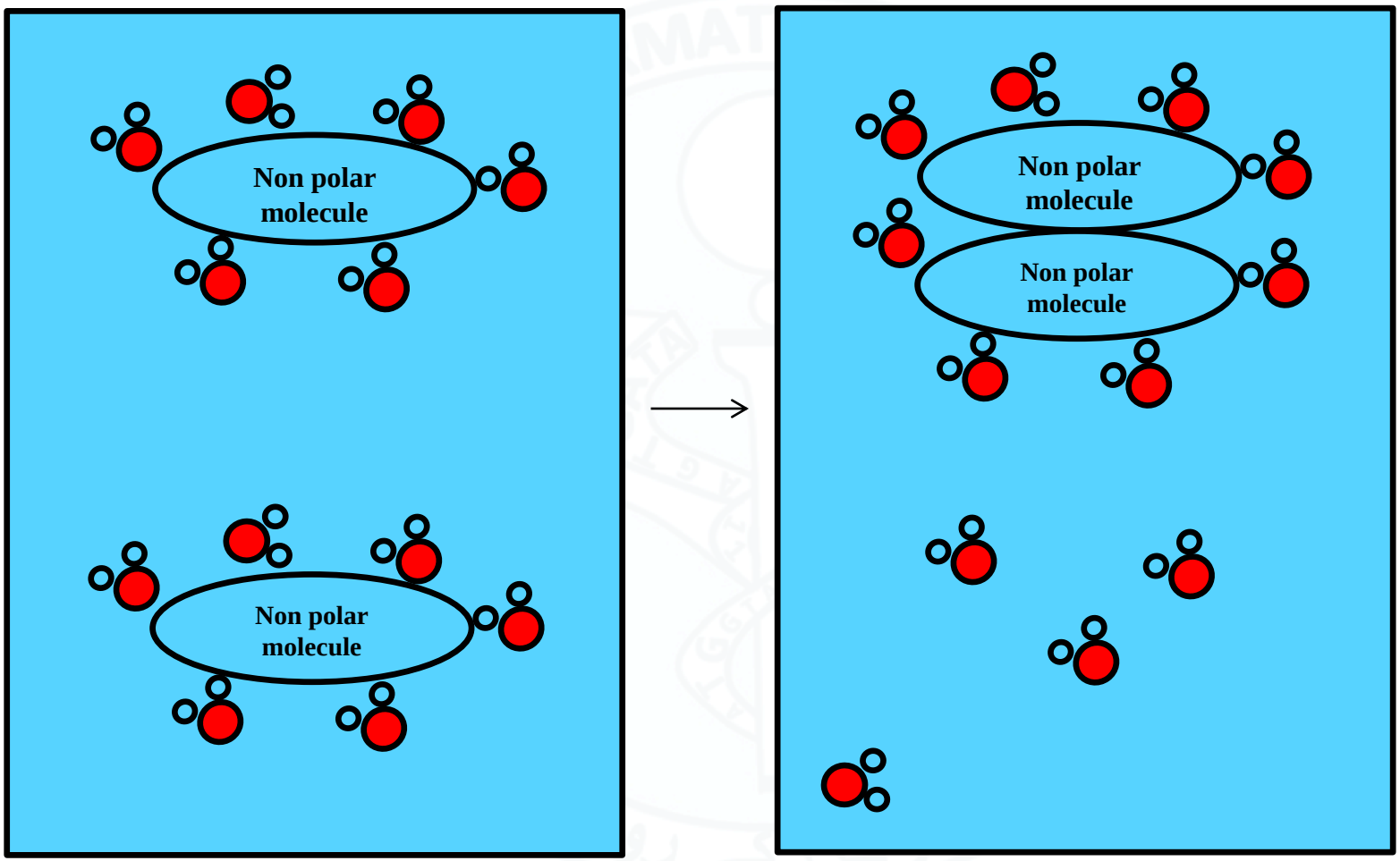
# Behavior of water around hydrophobic molecules

- When you add a drop of hydrophobic residues to water (oil drop concept)
  - Water forms an ordered cage around that



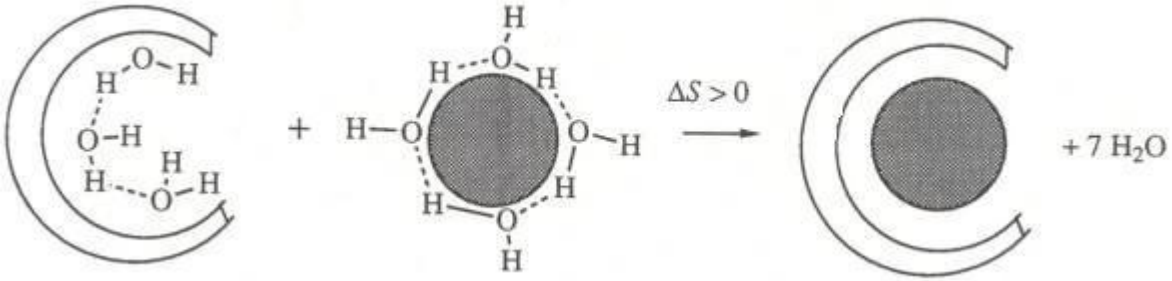
**Methane in Water Mimics the Hydration of Hydrophobic Surfaces**
Note that the water molecules avoid pointing their hydrogen-bonding groups toward the methane molecule. To do so would waste hydrogen bonds." In other words, they form and ordered cage around methane relative to the H-bonded waters in liquid water. The first-shell water molecules lose entropy to gain hydrogen bonding.

# The Hydrophobic Effect




Water molecules have less degrees of freedom in the clathrate cage arrangements because some H-bonds cannot point inside toward the hydrophobic sphere

# The Hydrophobic Effect

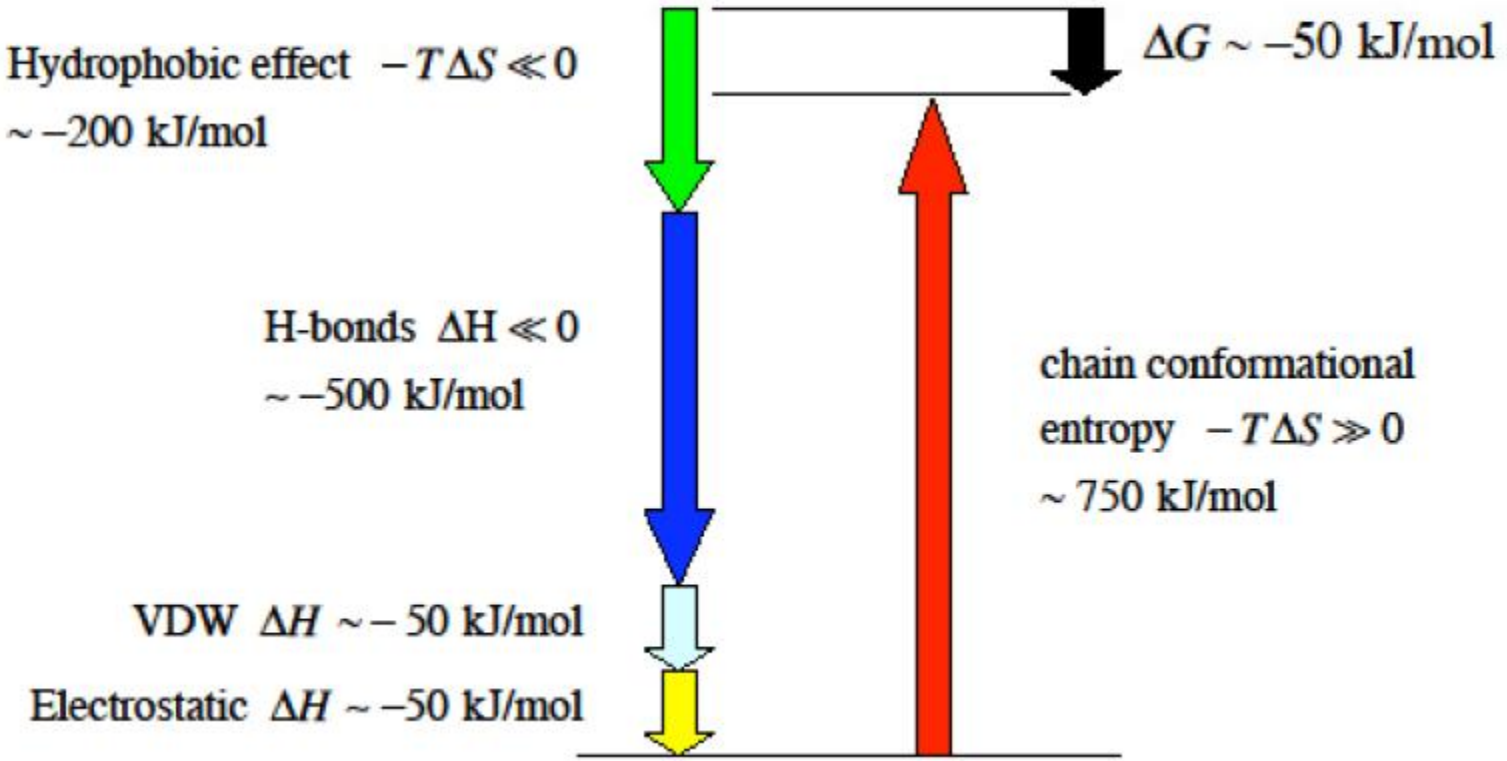



- There are “ordered water shells” around nonpolar groups and molecules in water. These hydration shells form to maximize H-bonding
- When nonpolar surfaces come in contact, the ordered water molecules are released into bulk solution, and the nonpolar surfaces are buried away from water in the process
- This leads to a large increase in ΔS, which drives association of the nonpolar surfaces

## Dissecting the free energy of protein folding

$$\text{Unfolded} \overset{\Delta G}{\rightleftharpoons} \text{Folded}$$

$$\Delta G = \Delta H - T\Delta S < 0, \quad \Delta G = \sim -50 \text{ kJ/mol}$$



# Energy Calculations

- Energy/Potential Function
  - Configuration of a protein (**X**)
    - For a sequence S of a protein, the location of the atoms in a protein with a given sequence can be determined if the torsion angles and rotamers are given.
  - For a given configuration of a protein, we can calculate its energy based on the interactions between the atoms by analyzing and combining the impact of different energy terms

# Things that can change in a protein

- Sequence
  - Sequence of the protein determines its native structure
  - The change in the sequence of the protein is likely to cause a change in the structure of the protein
  - Most of the time, a particular sequence gives rise to a certain structure

# Things that can change in a protein

- Torsion angles

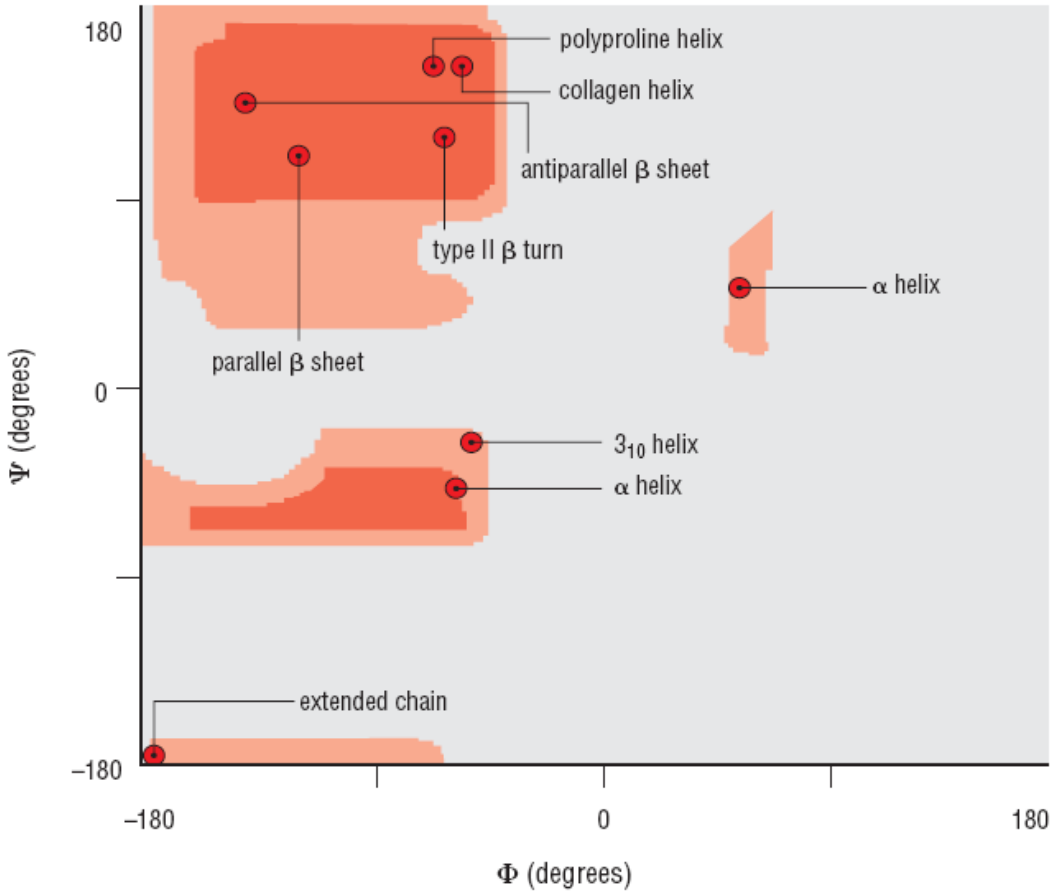
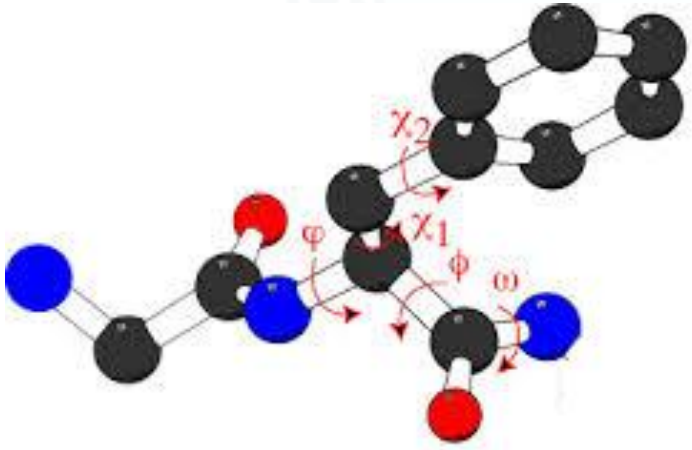
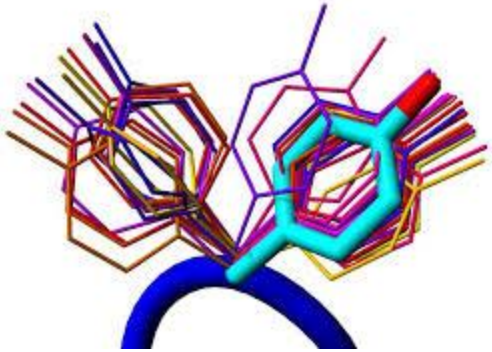



**Figure 1-11 Ramachandran plot** Shown in red are those combinations of the backbone torsion angles phi and psi (see Figure 1-9) that are "allowed" because they do not result in steric interference. The pink regions are allowed if some relaxation of steric hindrance is permitted. Common protein secondary structure elements are marked at the positions of their average phi, psi values. The isolated pink alpha-helical region on the right is actually for a left-handed helix, which is only rarely observed in short segments in proteins. The zero values of phi and psi are defined as the *trans* configuration.

# Side Chain Conformations

- The side chains are also rotatable
- Impacts the shape of the protein (and interactions)
- And the number of possible shapes



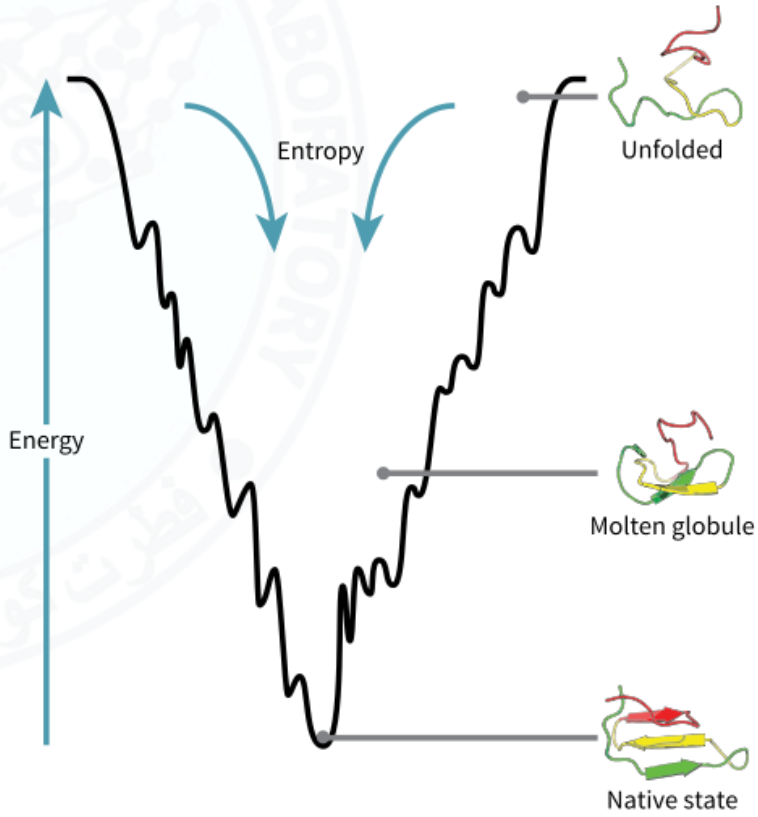



# Interpretation of the Energy Function

- Abstractly, the energy function can be written as
  - $E(S, \mathbf{X}; \boldsymbol{\theta})$
  - A mapping from the protein sequence and conformation spaces to its energy value
    - Parameters are denoted by $\boldsymbol{\theta}$
  - Given a sequence, the energy will be lowest for the native structure of the protein
- Physical interpretation
  - $\Delta \boldsymbol{G} = \Delta \boldsymbol{H} - \boldsymbol{T}\, \Delta \boldsymbol{S}$
  - Typically, the Entropy term is not explicitly modeled
  - A model of the energy of a protein

# Physical Interpreation

- Energy Landscape of Protein Folding
  - The native state of the protein is the conformation of the protein with the lowest energy




http://en.wikipedia.org/wiki/Folding_funnel

# Force Fields

- The energy function is a weighted summation of different contributing terms (electrostatics, VDW, …).
- These terms involve a set of parameters. The exact functional form of the energy function is called a “force field”.
- There are a number of different force fields available. For example:
  - AMBER, CHARMM, GROMOS, OPLS
- Example

$$V(r^N) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_a(\theta - \theta_0)^2$$

$$+ \sum_{\text{torsions}} \sum_{n} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[ \left( \frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

https://en.wikipedia.org/wiki/AMBER

# Points to Note

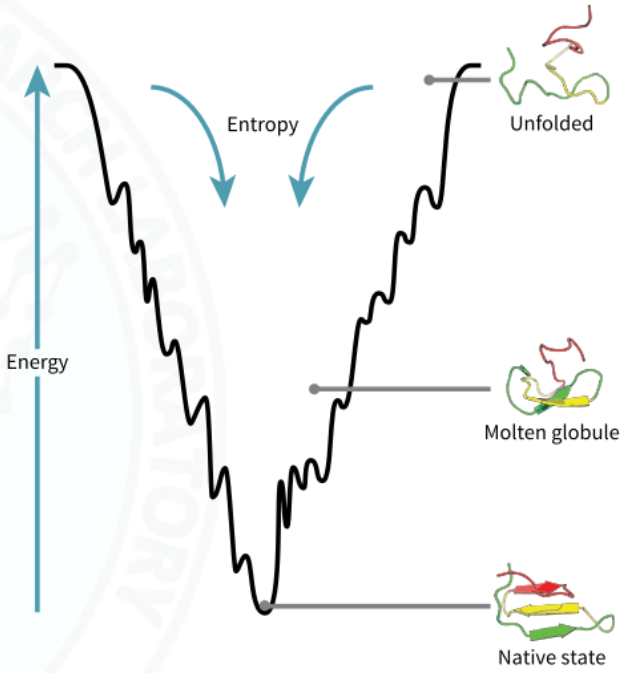
- This is only a model or an approximation.
  - Modeling molecular water
  - Empirical terms
  - Knowledge based
  - Quantum mechanical effects
    - http://www.sott.net/article/270402-Spooky-action-at-a-distance-Water-in-cells-behaves-in-complex-and-intricate-ways

- How to model what we cannot model?
  - Using machine learning aka statistical potentials
    - Given some training data, what are the chances of this particular conformation of the protein?
      - The higher the chances, the lower the energy and vice-versa

- Performing Energy Calculations
  - A number of software tools can calculate the energy of a protein
    - pyRosetta (http://www.pyrosetta.org/ )
    - SHARPEN
      - https://www.engr.colostate.edu/~cdasnow/snowlab_software.shtml

# What do these calculations help us do

- Protein Stability Calculations
  - Given: Structure(s) of a protein
  - Desired Output: An estimate of protein stability
    - The lower the energy of the protein the more stable it is. If two structures are given, we can calculate which one is likely to be more stable based on energy calculations
  - Computation
    - Simple Energy Calculations $E(S,\mathbf{X};\boldsymbol{\theta})$

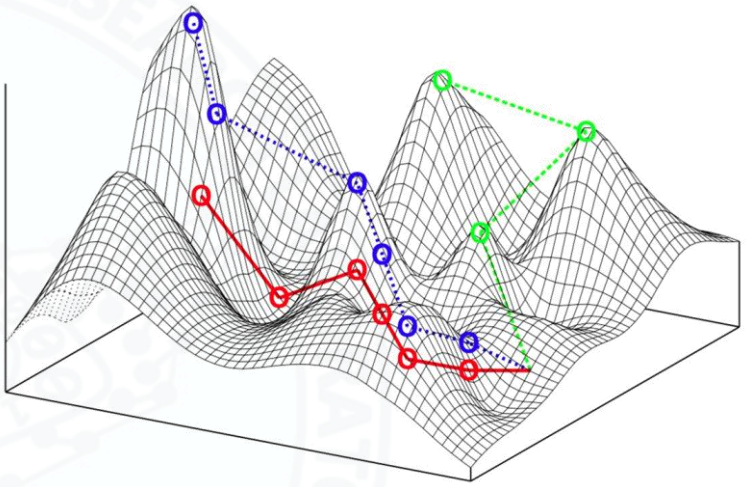



This is a consequence of the folding funnel model of the protein folding

# What do these calculations help us do

- Protein Structure Prediction
  - Given: The unfolded structure/sequence of a protein
  - Desired output: The native structure of a protein
    - The native structure of the protein is the conformation of the protein at the lowest energy.
  - Thus, we are interested in solving the following optimization problem:

$$X^* = argmin_X \, \mathrm{E}(S, X; \theta)$$



Structure space

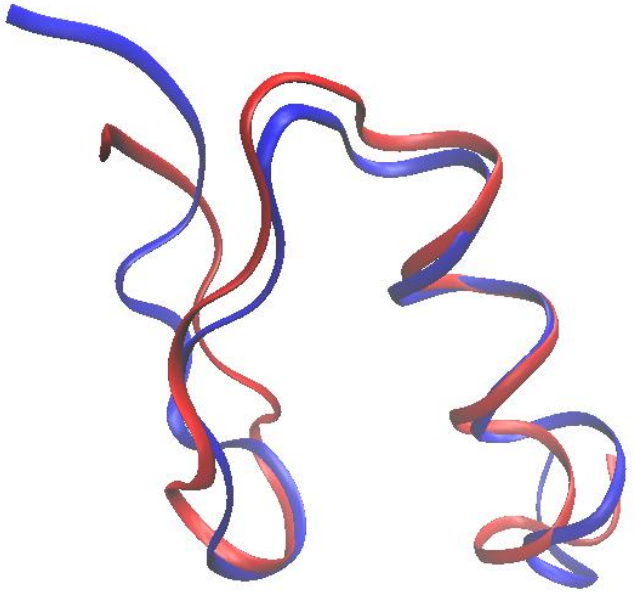
A change in, say, rotamer or torsion angles will cause the energy to change. This gives rise to the concept of a fitness landscape

https://en.wikipedia.org/wiki/Protein_structure_prediction

# What do these calculations help us do

- Protein Design / Inverse Folding
  - Given: The desired structure of a protein ($X_{desired}$)
  - Desired output: A sequence of the protein whose native structure is as close to the desired structure as possible
  - Objective: Design a protein which has a certain function. For that function, we sculpt a structure of the backbone and possibly of some side-chains involved in the desired function. Now, we would like to have a sequence that generates the desired structure. We use computing to obtain that sequence and then develop the protein chemically.
  - Mathematically,

$$S^* = argmin_S \, \mathrm{E}(S, X_{desired}; \theta)$$



FSD-1 (shown in blue, PDB id: 1FSV) was the first de novo computational design of a full protein.[2] The target fold was that of the zinc finger in residues 33-60 of the structure of protein Zif268 (shown in red, PDB id: 1ZAA). The designed sequence had very little sequence identity with any known protein sequence.

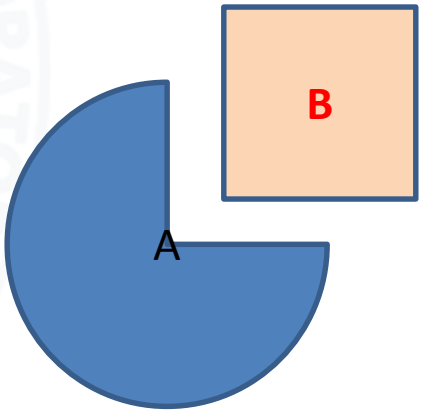
https://en.wikipedia.org/wiki/Protein_design

# What do these calculations help us do

- Protein Interactions Studies
  - Given: Two proteins A and B
  - Output:
    - Do these proteins interact
    - What parts of the two proteins interact
  - Two proteins will interact only if this leads to a decrease in the energy
    - Energy before binding:
      $$\Delta \boldsymbol{G}_A + \Delta \boldsymbol{G}_B$$
    - Energy before binding
      $$\Delta \boldsymbol{G}_{AB}$$
    - The proteins will bind only if the binding free energy:
      $\Delta\Delta \boldsymbol{G} = \Delta \boldsymbol{G}_{AB} - (\Delta \boldsymbol{G}_A + \Delta \boldsymbol{G}_B) < 0$
    - Stability of the interaction

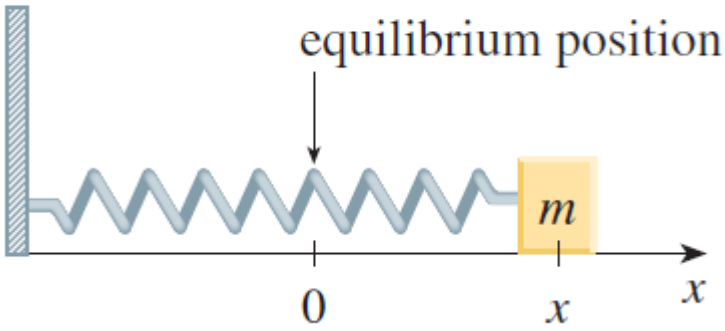
      Given by how negative the binding free energy is
    - Interaction Sites

      The two proteins will produce a joint conformation $X_{AB}$ which produces lowest $\Delta \boldsymbol{G}_{AB}$



# What do these calculations help us do

- Protein Dynamics
  - Given: Protein Structure
  - Desired output: How does this protein move or behave dynamically
  - If we know the energy function, the derivative of the energy function gives us the force. We can then use Newtonian mechanics to model the impact of that force on individual atoms to see how they move.




$$E = -kx^2$$

$$F = \frac{dE}{dx} = -kx$$

$$m\frac{d^2x}{dt^2} = -kx \quad \text{or} \quad m\frac{d^2x}{dt^2} + kx = 0$$

$$x(t) = A\cos(\omega t + \delta)$$

A simple example of dynamics simulation based on energy calculations.

https://en.wikipedia.org/wiki/Molecular_dynamics

# Protein Dynamics

$$F_i = -\nabla_i E$$

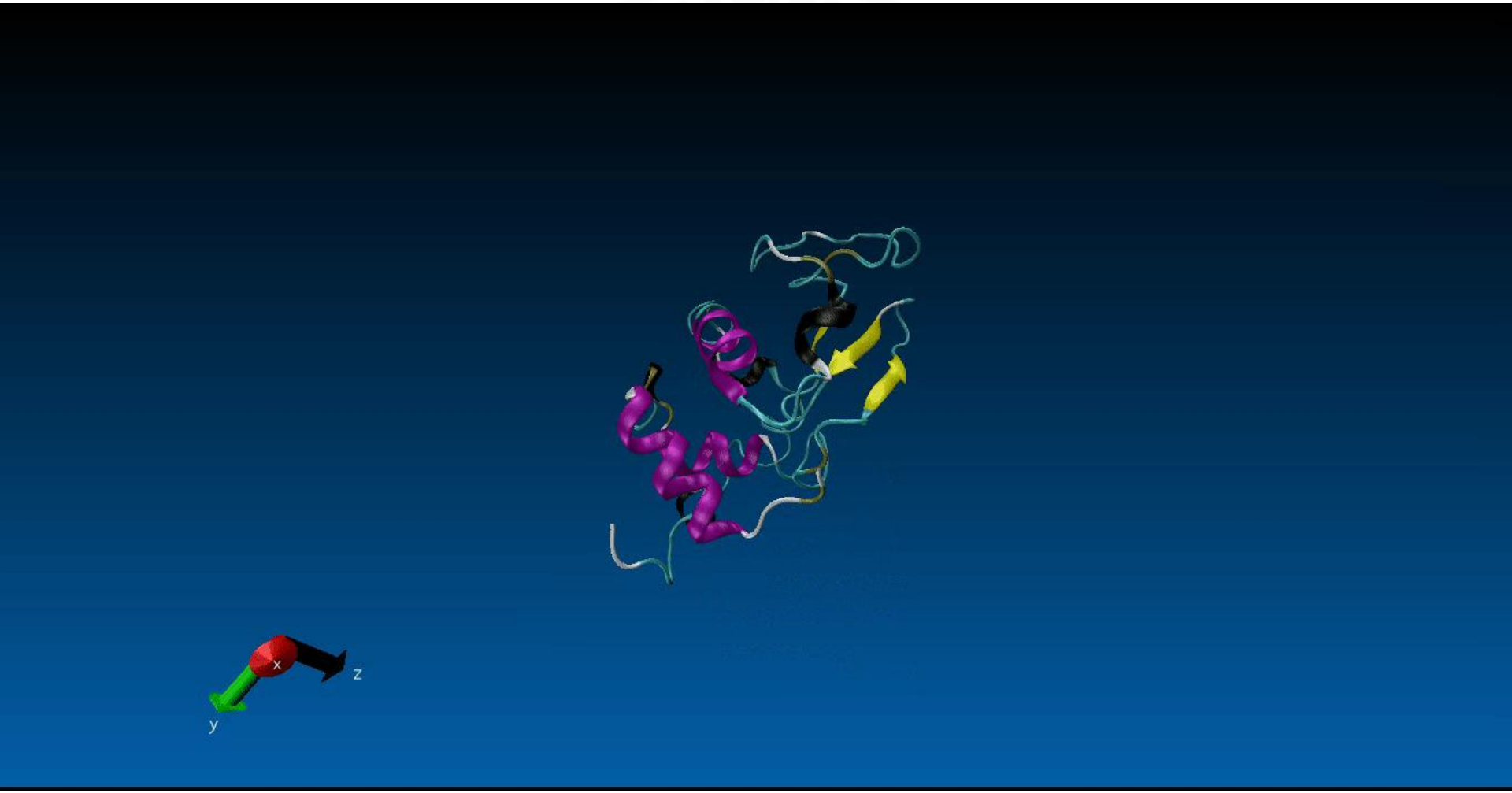
$$F_i = m_i a_i$$

$$-\frac{dE}{dr_i} = m_i \frac{d^2 r_i}{dt^2}$$

E is the potential as described by the force field

To calculate a trajectory of motion of an atom, one only needs the initial positions of the atoms, an initial distribution of velocities and the acceleration, which is determined by the gradient of the potential energy function. The equations of motion are deterministic, e.g., the positions and the velocities at time zero determine the positions and velocities at all other times, t. The initial positions can be obtained from experimental structures, such as the x-ray crystal structure of the protein or the solution structure determined by NMR spectroscopy.

# 10ns Molecular Dynamics Simulation of a protein (1AKL) with GROMACS



# Protein Folding Simulation Video


# Protein Structure Prediction

**Dr. Fayyaz ul Amir Afsar Minhas**
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

# Tertiary Structure Prediction Methods

- Comparative/Homology Modeling Based Techniques
- Fold Recognition and Threading
- De novo prediction (from first principles)

- According to Science, the problem remains one of the top 125 outstanding issues in modern science.[1]
- The protein folding problem: when will it be solved?

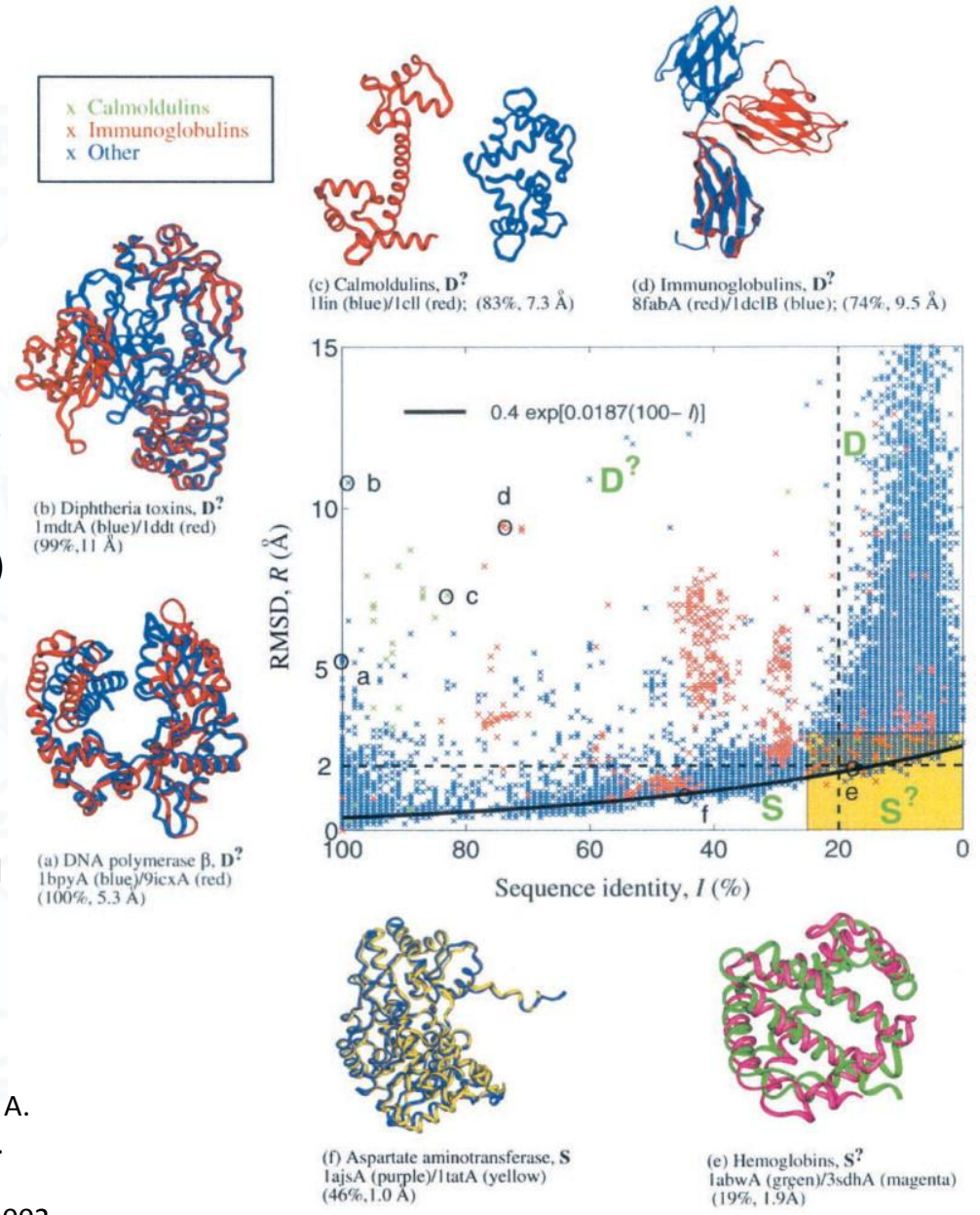
"So Much More to Know …," Science, vol. 309, no. 5731, pp. 78–102, Jul. 2005.

# Protein Tertiary Structure Prediction

- Sequence to Structure Relationship
- Protein Energetics

# Seq. vs. RMSD

- Sequence/structure map of 53,383 protein pairs displayed as a root-mean-square deviation (RMSD or *R*) versus percent sequence identity, *I*, plot. The 465 probe proteins listed in Table 1 were used to generate these pairs using the FSSP algorithm. Only alignments >75 residues are included. Sequence/structure coordinates are marked by blue × symbols except for those resulting from calmodulin (1cll, *green*) and immunoglobulin (8fabA, *red*) probes. The map is subdivided into regions as discussed in the text: expected similarity (S); unexpected similarity ($S^?$); expected dissimilarity (D); and unexpected dissimilarity ($D^?$). Superpositions of selected protein pairs in different regions are marked and displayed. The thick black line corresponds to the empirical exponential function of Eq. 2.




H. Hark Gan, R. A. Perlow, S. Roy, J. Ko, M. Wu, J. Huang, S. Yan, A. Nicoletta, J. Vafai, D. Sun, L. Wang, J. E. Noah, S. Pasquali, and T. Schlick, "Analysis of Protein Sequence/Structure Similarity Relationships," *Biophys. J.*, vol. 83, no. 5, pp. 2781–2791, Nov. 2002.

# Sequence vs. Structure

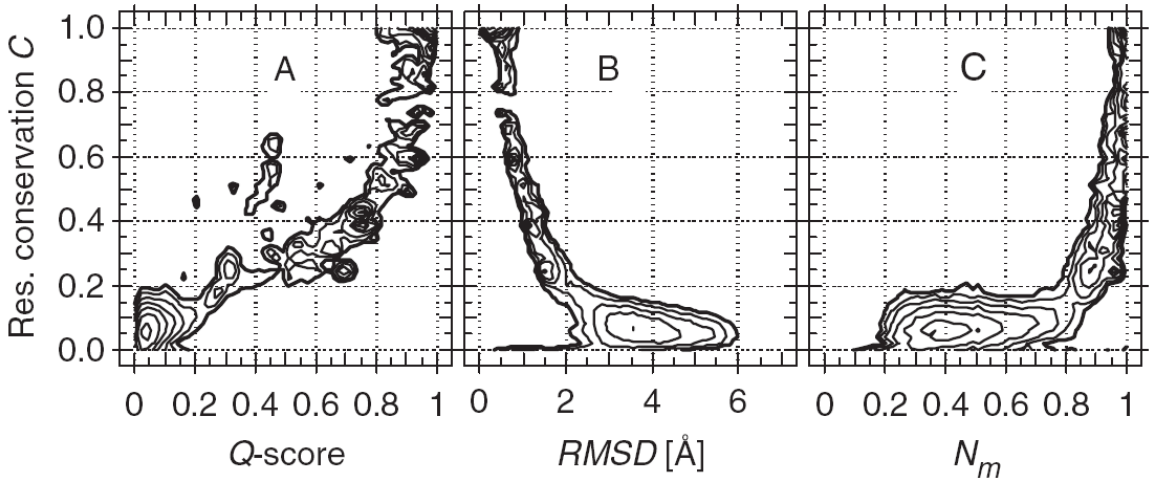



**Fig. 1.** Correlations between residue conservation $C$ and structure similarity scores: (A) $Q$-score (B) r.m.s.d. (C) normalized alignment length $N_m$ [cf. Equation (1)], represented as contour maps of reduced density of probability to find the corresponding pairs of similar structures in the PDB [cf. Equation (2)], from Krissinel and Henrick, (2004). The outermost contours correspond to the level of 0.05 from the maximum. The data suggest that, on average, structures are dissimilar at $C \leq C_0 = 0.2$, see discussion in the text.

E. Krissinel, "On the relationship between sequence and structure similarities in proteomics," *Bioinformatics*, vol. 23, no. 6, pp. 717–723, Mar. 2007.

# Comparative Modeling

- Principle
  - Sequence drives structure
    - Evolutionarily related sequences exhibit similar three dimensional structure

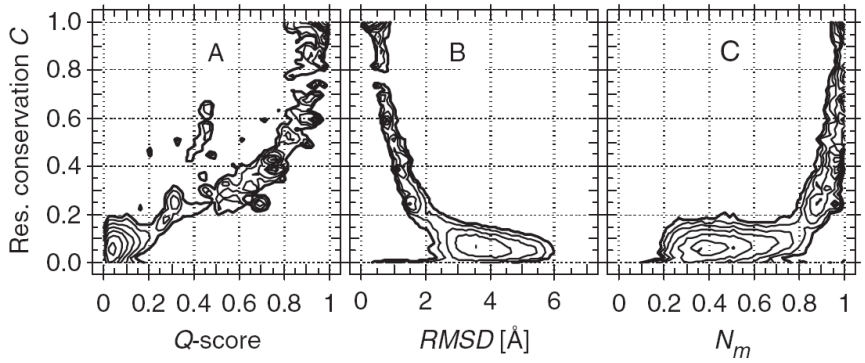



**Fig. 1.** Correlations between residue conservation $C$ and structure similarity scores: (A) $Q$-score (B) r.m.s.d. (C) normalized alignment length $N_m$ [cf. Equation (1)], represented as contour maps of reduced density of probability to find the corresponding pairs of similar structures in the PDB [cf. Equation (2)], from Krissinel and Henrick, (2004). The outermost contours correspond to the level of 0.05 from the maximum. The data suggest that, on average, structures are dissimilar at $C \leq C_0 = 0.2$, see discussion in the text.

# H. Modeling

- Steps
  - Homology detection
  - Alignment of target sequence to template structures
  - Modeling of structurally conserved regions using known templates
  - Modeling side chains and loops which are different from the templates
  - Refinement

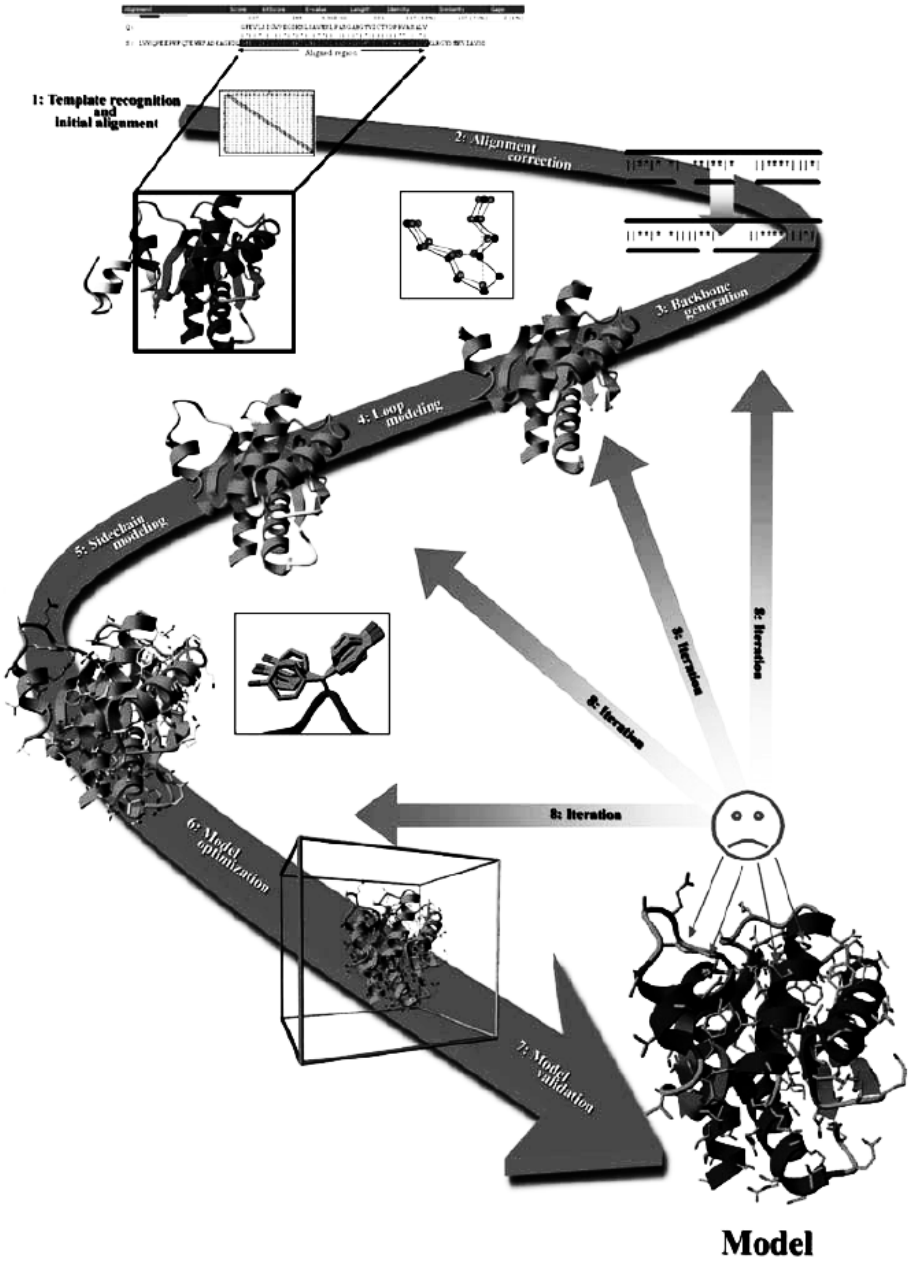



Figure 30.4. The process of building a ''model'' by homology to a ''template.'' The numbers in the plot correspond to the step numbers in the subsequent section. (Colored version of this Figure is available for viewing at http://swift.cmbi.ru.nl/material/)

(READ)
Venselaar et al., "Homology Modeling", Chapter 30 in Structural Bioinformatics, editors: Gu and Bourne, 2011.

# Comparative Modeling

- During alignment or homology detection, should we use protein sequences or the DNA sequence that produced the protein?
  - Why?
- If two proteins have >30% sequence identity
  - More than 80% of the C-alpha atoms can be expected to be within 3.5 Angsrtoms of their true positions
  - Otherwise errors

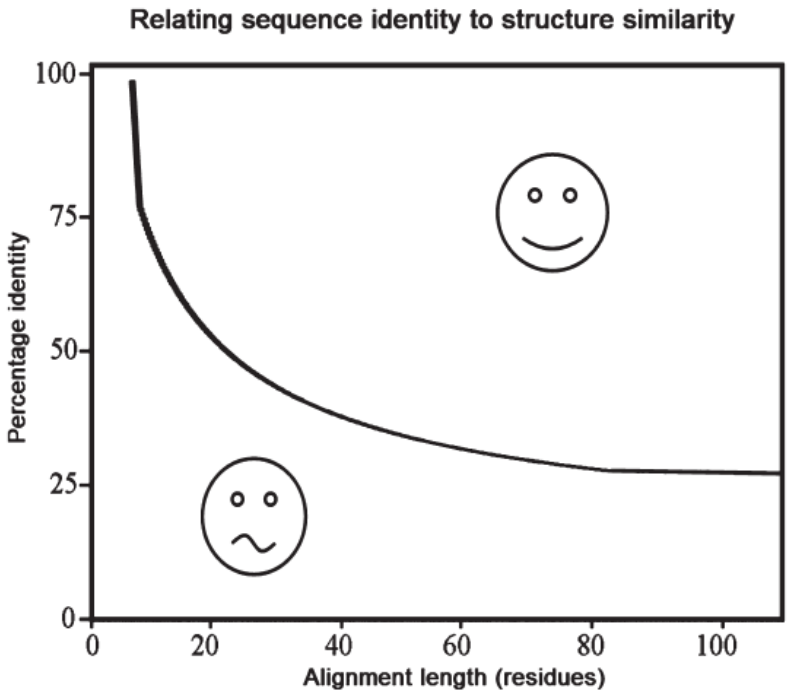



**Figure 2.** HSSP sequence alignment threshold. The structural homology plot (18) describes at which percentage sequence identity an alignment of a given length is an indication that the aligned proteins have a similar structure. The dark curve gives the cut-off below which no structural similarity can be inferred. This is frequently used in the context of homology modelling: alignments above the curve indicate that it is possible to make a fairly reliable homology model from the aligned template; alignments below the curve mean that a homology model should be handled with care.

B. Rost, "Twilight zone of protein sequence alignments," *Protein Eng.*, vol. 12, no. 2, pp. 85–94, Feb. 1999.

# Examples

TABLE 30.1 A Few Examples of the Online Available Homology Modeling Servers

| Server Name | URL |
|---|---|
| Automatic Homology Modeling Servers | |
| 3D-Jigsaw | http://www.bmm.icnet.uk/servers/3djigsaw/ |
| CPHModels | http://www.cbs.dtu.dk/services/CPHmodels/ |
| EsyPred3D | http://www.fundp.ac.be/urbm/bioinfo/esypred/ |
| Robetta | http://robetta.bakerlab.org/ |
| SwissModel | http://swissmodel.expasy.org/ |
| TASSER-lite | http://cssb.biology.gatech.edu/skolnick/webservice/tasserlite/index.html |
| Semiautomatically Homology Modeling Servers | |
| HOMER | http://protein.cribi.unipd.it/homer/help.html |
| WHAT If | http://swift.cmbi.kun.nl/WIWWWI/ |

# SWISS-model

- Use BLAST to fetch homologous sequences
- If no suitable templates are found, HHSearch is used for detection of remotely related sequences
  - Database: Uses SWISS-MODEL Template Library which is derived from PDB
- Alignment of target sequence and template structure(s)
- Model Building and Energy Minimization
  - A rigid fragment assembly approach for modeling
- Assessment of model quality
  - QMEAN, ANOLEA, GROMOS

http://en.wikipedia.org/wiki/Swiss-model

# Fold Recognition / Threading

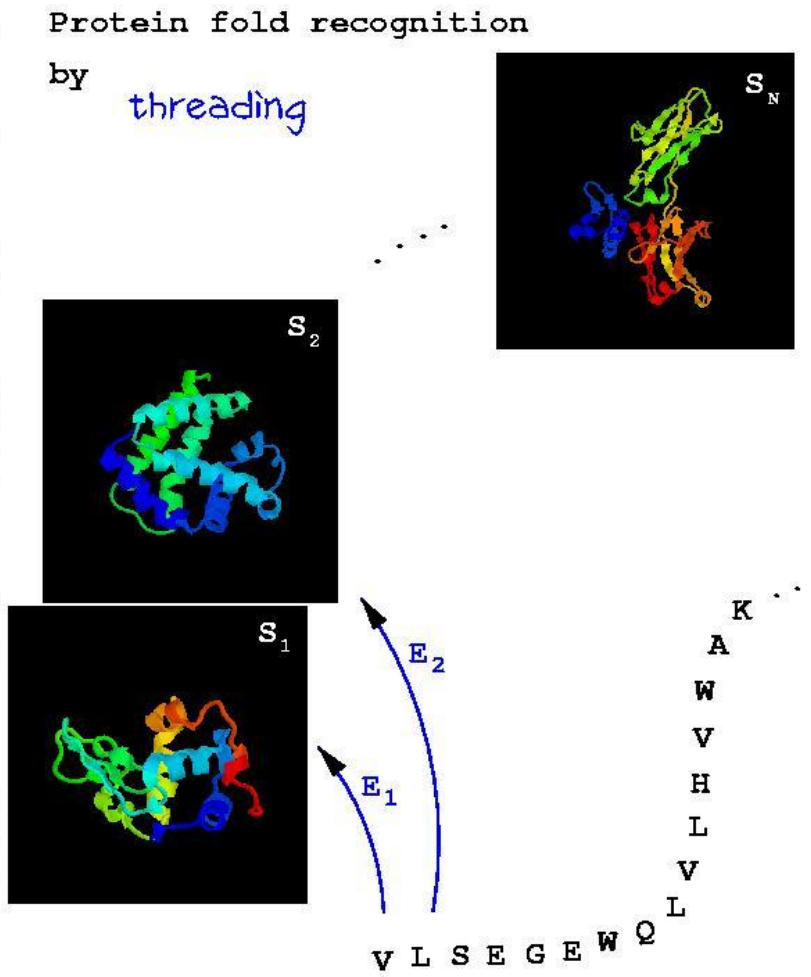
- Main idea:
  - The number of unique folds is limited

- Fit a target sequence to a known structure in a library of folds
  - Inverse folding
  - threading

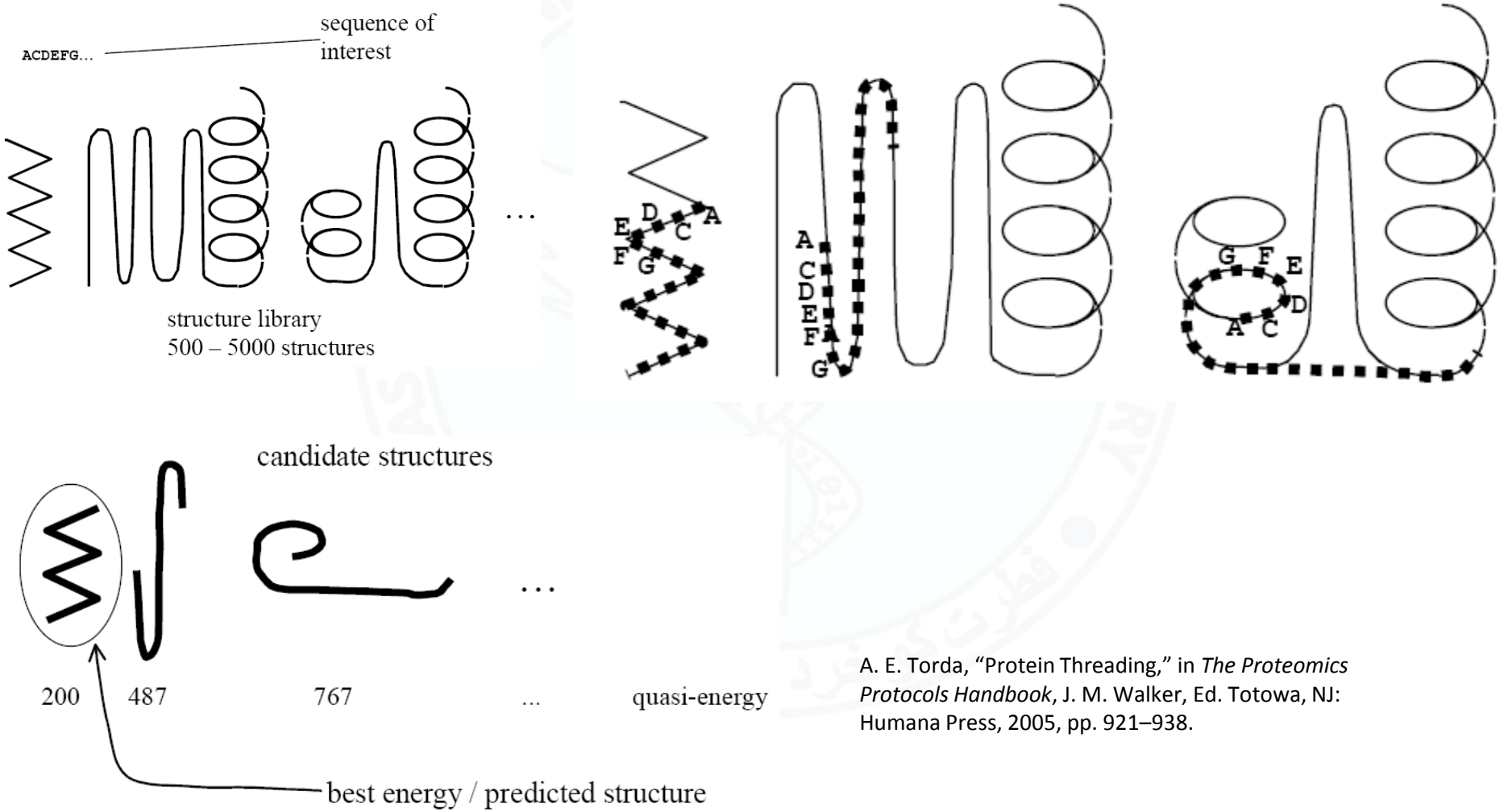
http://en.wikipedia.org/wiki/Threading_(protein_sequence)

# Threading

- Components
  - Library of templates (SCOP, FSSP…)
  - Scoring Function
    - Measures the fitness of a sequence to structure alignment
    - Abstract form: F(sequence, structure)
      - F(sequence, structure) is low if there is good correspondence between the sequence and the given structure
      - High otherwise
  - Sequence to structure alignment
    - Find the best alignment between a fixed sequence (given) and structure



# Threading overview




A. E. Torda, "Protein Threading," in *The Proteomics Protocols Handbook*, J. M. Walker, Ed. Totowa, NJ: Humana Press, 2005, pp. 921–938.

# Threading

- Scoring Function should consider
  - Environmental preferences
  - Secondary structure preferences
  - Solvent exposure
  - Pairwise interactions with neighboring amino acids
- Knowledge based potentials
  - Example: How likely is it that two residues are a certain distance apart?

# Sequence to Structure Alignment

- Find the best alignment of the given sequence to the template structures
- Use the scoring function together with a search technique
  - Monte-Carlo
  - Simulated Annealing
  - Dynamic Programming

# Fold Recognition

- Now we have a number of best alignments of templates to the given sequence
  - Which one should we use?
    - The best one?
    - Which one do we want?
      - The one that occurs in nature
  - Use machine learning to weed out the wrong ones

# Methods

- Hhpred
- RAPTORX
- Phyre

- Limitations
  - A completely new fold?

http://en.wikipedia.org/wiki/Threading_%28protein_sequence%29

# De novo protein structure prediction

- With database support
  - Rosetta
  - Knowledge based scoring function
- Without database support
  - Using first principles without any database information
  - Physics based scoring function

J. Lee, S. Wu, and Y. Zhang, "Ab Initio Protein Structure Prediction," in *From Protein Structure to Function with Bioinformatics*, D. J. Rigden, Ed. Springer Netherlands, 2009, pp. 3–25.

# Rosetta

- Fragment Assembly Approach

- Uses PDB information to estimate the possible conformations for local sequence segments
  - 3 and 9 residue matches between the query sequence and a given structure
- Fragment substitution
- Initial low resolution refinement using a scoring function
- Second stage refinement

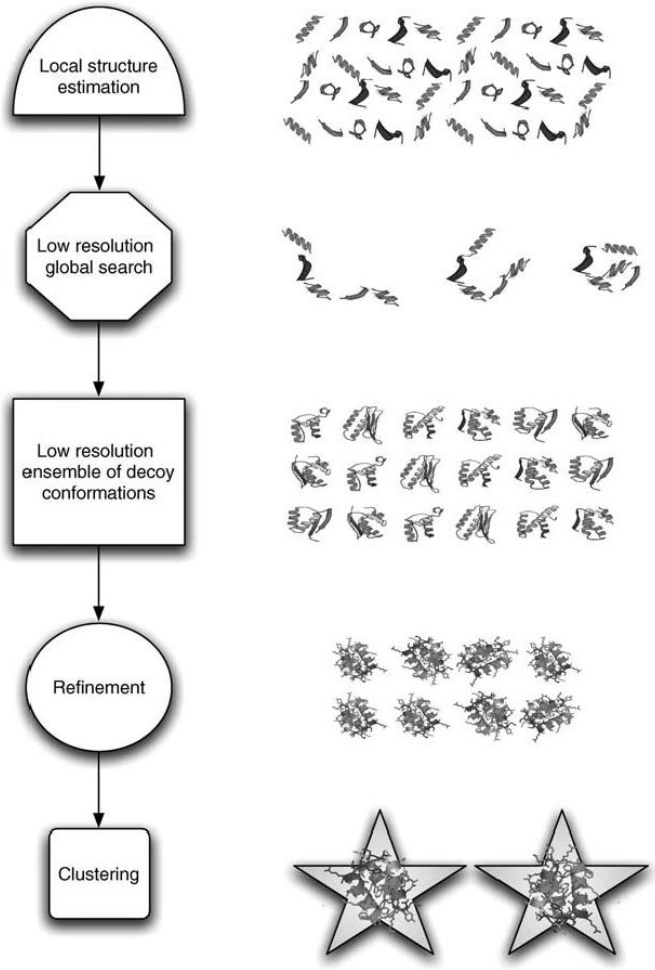



Figure 32.2. Rosetta structure prediction protocol: Rosetta begins by determining local structure conformations (fragments) for 3- and 9-mer stretches of the input sequence. Then multiple fragment substitution simulated annealing searches are done to find the best arrangement of the fragments according to Rosetta's low-resolution scoring function (Table 32.1). The resulting structures then undergo a high-resolution refinement step based on a more physically based scoring function (Table 32.2). Finally, the structures are clustered by RMSD (to each other, as the native is unknown) and the centers of the largest clusters are chosen as representative folds (the centers of the largest clusters are likely to be correct fold predictions).

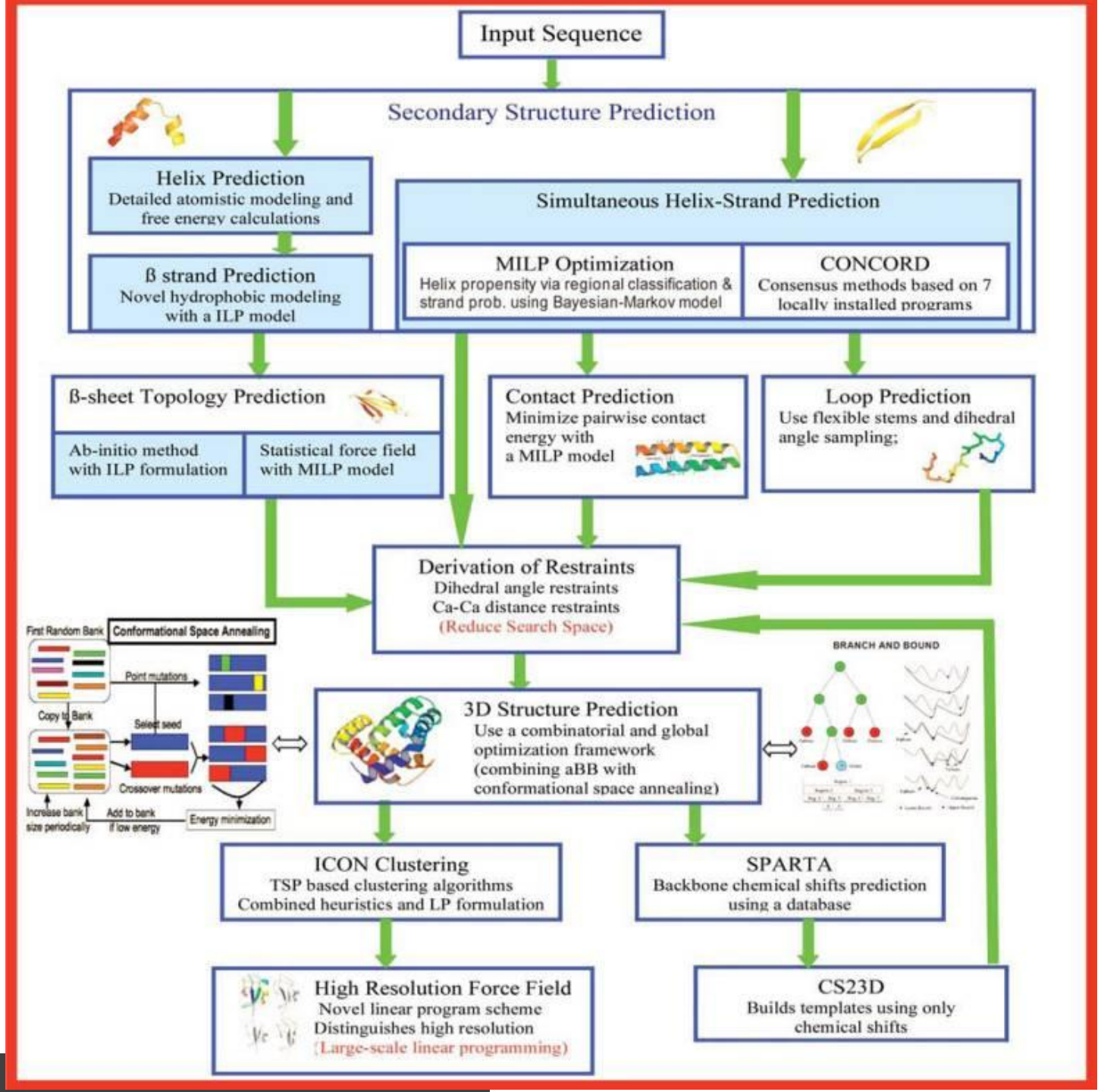
http://en.wikipedia.org/wiki/Rosetta@home

# First principles without database information

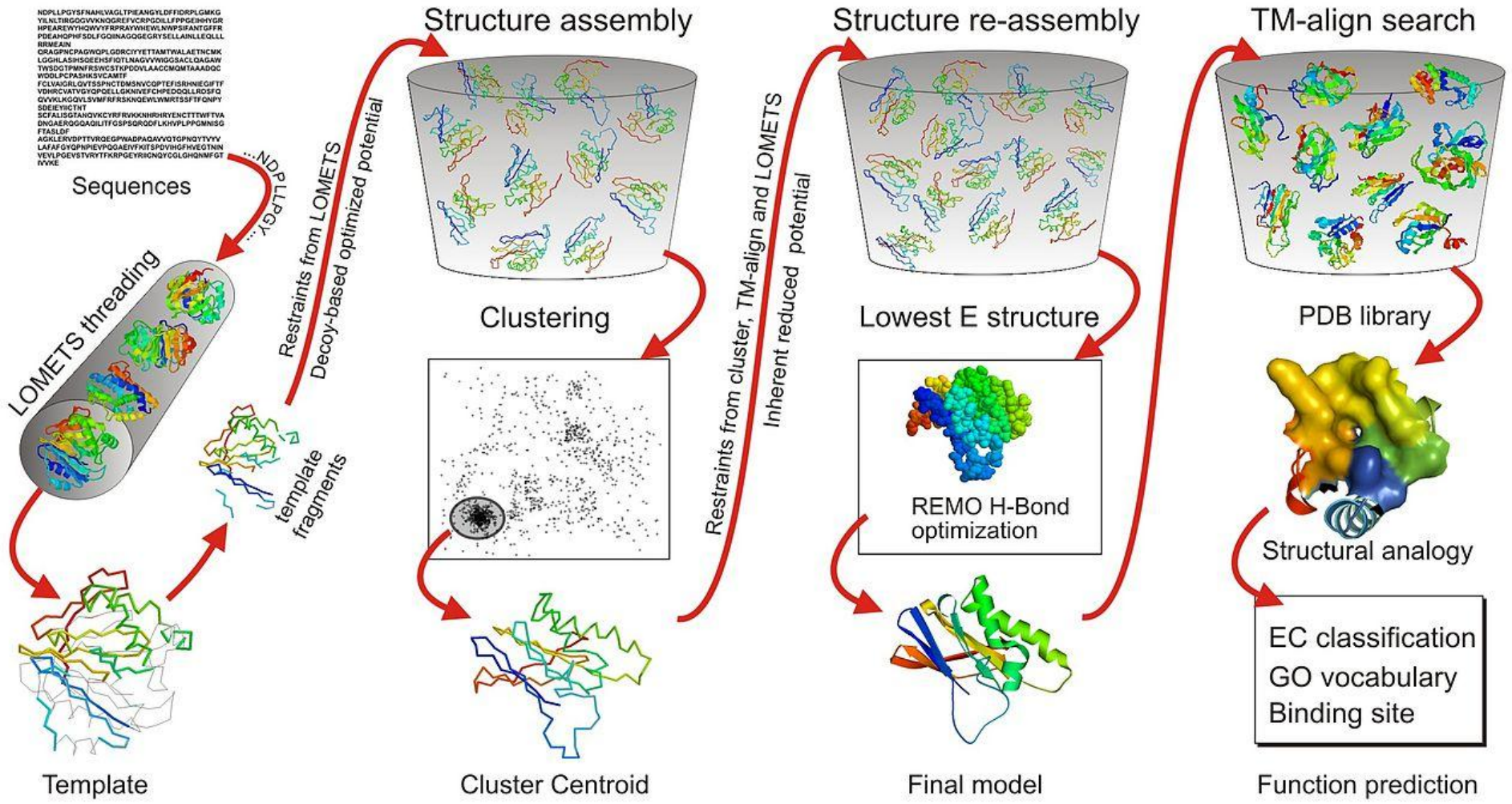
- Find the lowest free energy state of a protein using only physics laws and the amino acid sequence

- Use a force field and optimize it across conformations

# ASTRO FOLD


Input Sequence
Secondary Structure Prediction
Helix Prediction
Detailed atomistic modeling and free energy calculations
ß strand Prediction
Novel hydrophobic modeling with a ILP model
Simultaneous Helix-Strand Prediction
MILP Optimization
Helix propensity via regional classification & strand prob. using Bayesian-Markov model
CONCORD
Consensus methods based on 7 locally installed programs
ß-sheet Topology Prediction
Ab-initio method with ILP formulation
Statistical force field with MILP model
Contact Prediction
Minimize pairwise contact energy with a MILP model
Loop Prediction
Use flexible stems and dihedral angle sampling;
Derivation of Restraints
Dihedral angle restraints
Ca-Ca distance restraints
(Reduce Search Space)
First Random Bank
Conformational Space Annealing
Point mutations
Copy to Bank
Select seed
Crossover mutations
Increase bank size periodically
Add to bank if low energy
Energy minimization
3D Structure Prediction
Use a combinatorial and global optimization framework (combining aBB with conformational space annealing)
BRANCH AND BOUND
ICON Clustering
TSP based clustering algorithms
Combined heuristics and LP formulation
SPARTA
Backbone chemical shifts prediction using a database
High Resolution Force Field
Novel linear program scheme
Distinguishes high resolution
(Large-scale linear programming)
CS23D
Builds templates using only chemical shifts

# I-TASSER


Sequences
LOMETS threading
template fragments
Template
Restraints from LOMETS
Decoy-based optimized potential
Structure assembly
Clustering
Cluster Centroid
Restraints from cluster, TM-align and LOMETS
Inherent reduced potential
Structure re-assembly
Lowest E structure
REMO H-Bond optimization
Final model
TM-align search
PDB library
Structural analogy
EC classification
GO vocabulary
Binding site
Function prediction

# Comparison

- CASP
- CAMEO
  - http://www.cameo3d.org/sp/1-year/

# Determining the Accessible Surface Area

- Accessible surface area
  - Relative ASA
- Role a ball around the molecule to obtain the surface
- Methods:
  - STRIDE

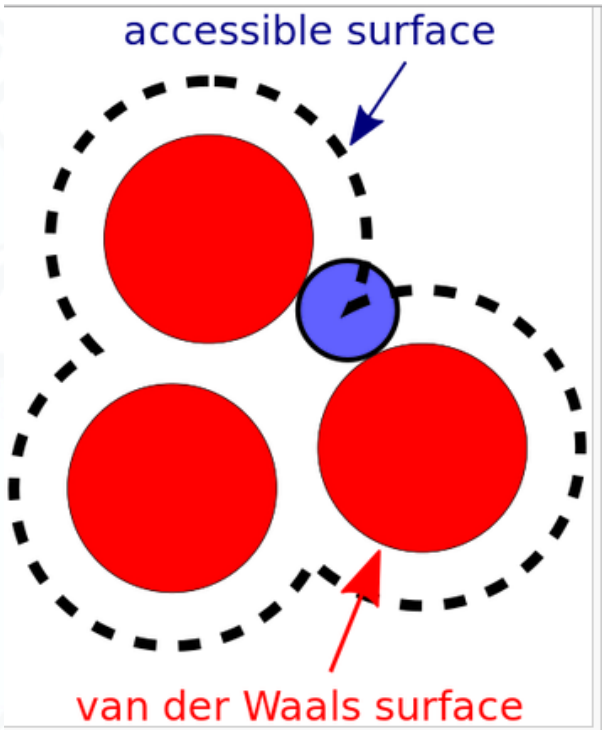



Illustration of the solvent accessible surface in comparison to the van der Waals surface. The van der Waals surface as given by the atomic radii is shown in red. The accessible surface is drawn with dashed lines and is created by tracing the center of the probe sphere (in blue) as it rolls along the van der Waals surface. Note that the probe radius depicted here is of smaller scale than the typical 1.4Å.

# Computational Prediction of Secondary and Supersecondary structures

- Chen, Ke, and Lukasz Kurgan. 2013. "Computational Prediction of Secondary and Supersecondary Structures." In *Protein Supersecondary Structures*, edited by Alexander E. Kister, 63–86. Methods in Molecular Biology 932. Humana Press. http://dx.doi.org/10.1007/978-1-62703-065-6_5.

# End of Lecture