Hydrophobicity scales

Kyte-Doolittle

Alanine	1.8
Arginine	-4.5
Asparagine	-3.5
Aspartic acid	-3.5
Cysteine	2.5
Glutamine	-3.5
Glutamic acid	-3.5
Glycine	-0.4
Histidine	-3.2
Isoleucine	4.5
Leucine	3.8
Lysine	-3.9
Methionine	1.9
Phenylalanine	2.8
Proline	-1.6
Serine	-0.8
Threonine	-0.7
Tryptophan	-0.9
Tyrosine	-1.3
Valine	4.2

A positive value indicates a hydrophobic residue and a negative value a hydrophilic residue

Hydropathy index

Hydropathy plots

Sliding Window Approach

Calculate property for first sub-sequence

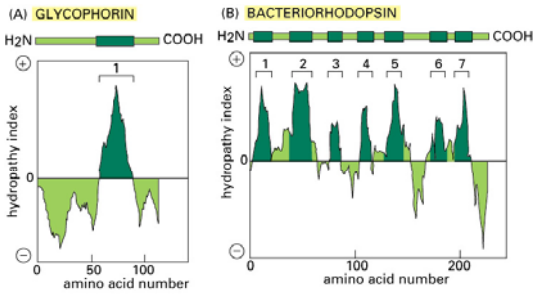
```
I L I K E I R
 4.50+3.80+4.50-3.90
-3.50+4.50-4.50 = 5.40

= 5.4/7=0.77
```

Move to the next position

Kyte-Doolittle

Alanine	1.8
Arginine	-4.5
Asparagine	-3.5
Aspartic acid	-3.5
Cysteine	2.5
Glutamine	-3.5
Glutamic acid	-3.5
Glycine	-0.4
Histidine	-3.2
Isoleucine	4.5
Leucine	3.8
Lysine	-3.9
Methionine	1.9
Phenylalanine	2.8
Proline	-1.6
Serine	-0.8
Threonine	-0.7
Tryptophan	-0.9
Tyrosine	-1.3
Valine	4.2



Methods for structure prediction

1.) Prediction of secondary structure
- a. method of Chou & Fassman
- b. neural networks

} predictive methods

2.) Prediction of tertiary structure
- a. *ab initio* structure prediction
- b. threading
  - -1D-3D profiles
  - - knowledge based potentials
- c. homology modelling

} modelling methods

Secondary Structure prediction

Three-state model: helix, strand, coil

Given a protein sequence:

NWVLSTAADMQGVVT DGMASGLDKD...

Predict a secondary structure sequence:

LLEEEELLLLHHHHHH HHHHLHHHL...

Chou-Fasman Parameters

Name	Abbrv	P(a)	P(b)	P(turn)
Alanine	A	142	83	66
Arginine	R	98	93	95
Aspartic Acid	D	101	54	146
Asparagine	N	67	89	156
Cysteine	C	70	119	119
Glutamic Acid	E	151	37	74
Glutamine	Q	111	110	98
Glycine	G	57	75	156
Histidine	H	100	87	95
Isoleucine	I	108	160	47
Leucine	L	121	130	59
Lysine	K	114	74	101
Methionine	M	145	105	60
Phenylalanine	F	113	138	60
Proline	P	57	55	152
Serine	S	77	75	143
Threonine	T	83	119	96
Tryptophan	W	108	137	96
Tyrosine	Y	69	147	114
Valine	V	106	170	50

Chou-Fasman Algorithm

- Identify α-helices
  - 4 out of 6 contiguous amino acids that have P(a) > 100
  - Extend the region until 4 amino acids with P(a) < 100 found
  - Compute ΣP(a) and ΣP(b); If the region is >5 residues and ΣP(a) > ΣP(b) *identify as a helix*
- Repeat for β-sheets [use P(b)]
- If an α and a β region overlap, the overlapping region is predicted according to ΣP(a) and ΣP(b)

Remember

helix - 4 out of 6 residues with high helix propensity (P > 100)
sheet - 3 out of 5 residues with high sheet propensity (P > 100)

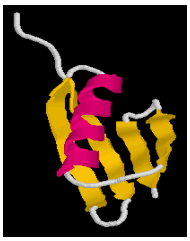
Name	Abbrv	P(a)	P(b)	P(turn)
Alanine	A	142	83	66
Arginine	R	98	93	95
Aspartic Acid	D	101	54	146
Asparagine	N	67	89	156
Cysteine	C	70	119	119
Glutamic Acid	E	151	37	74
Glutamine	Q	111	110	98
Glycine	G	57	75	156
Histidine	H	100	87	95
Isoleucine	I	108	160	47
Leucine	L	121	130	59
Lysine	K	114	74	101
Methionine	M	145	105	60
Phenylalanine	F	113	138	60
Proline	P	57	55	152
Serine	S	77	75	143
Threonine	T	83	119	96
Tryptophan	W	108	137	96
Tyrosine	Y	69	147	114
Valine	V	106	170	50

	T	S	P	T	A	E	L	M	R	S	T	G
P(H)	69	77	57	69	142	151	121	145	98	77	69	57

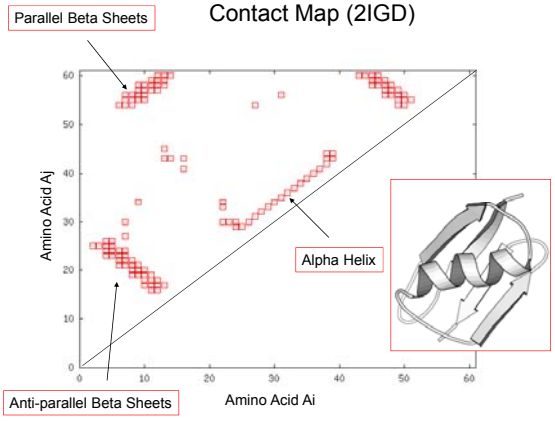
	T	S	P	T	A	E	L	M	R	S	T	G
P(H)	69	77	57	69	142	151	121	145	98	77	69	57

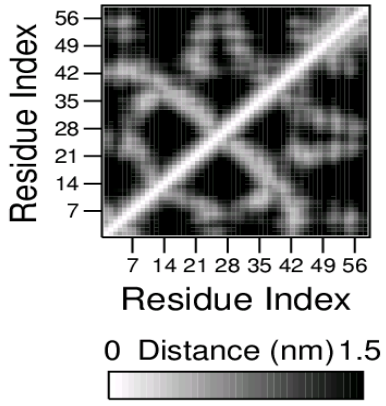
Contact Map

- Amino acids Ai and Aj are in contact if their 3D distance is less than a *contact threshold* (e.g., 7 Angstroms)
- Sequence separation is given as |i-j|
- Contact map C is a symmetric N x N matrix with
  - C(i,j) = 1 if Ai and Aj are in contact
  - C(i,j) = 0 otherwise
- Consider all pairs with |i-j| >= 4



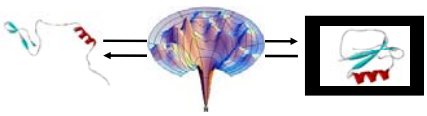
Contact Map (2IGD)





Features of the native state

-well defined 3D structure
-Isoelectric point (pI)
-Some characterized molecular function



- Many proteins fold spontaneously to their native structure
- Protein folding is relatively fast
- Chaperones speed up folding, but do not alter the structure

Forces driving protein folding

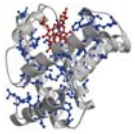
- It is believed that *hydrophobic collapse* is a key driving force for protein folding
  - Hydrophobic core
  - Polar surface interacting with solvent
- Minimum volume (no cavities)
- Disulfide bond formation stabilizes
- Hydrogen bonds
- Polar and electrostatic interactions

Native state is typically only 5 to 10 kcal/mole more stable than the unfolded form

The Protein Folding Problem

Levinthal's paradox – Consider a 100 residue protein.
If each residue can take only 3 positions,
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!

MACGT... ? 

"Given a particular sequence of amino acid residues (primary structure), what will the tertiary/quaternary structure of the resulting protein be?"

Four models that could account for the rapid rate of protein folding during biological protein synthesis.

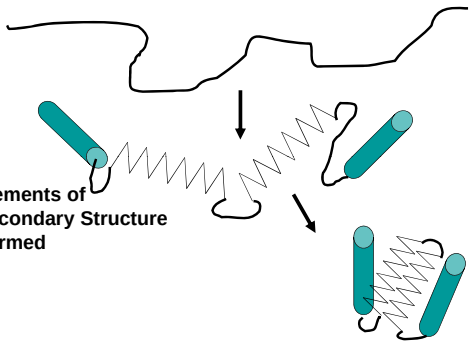
- The Framework Model

- The Nucleation Model

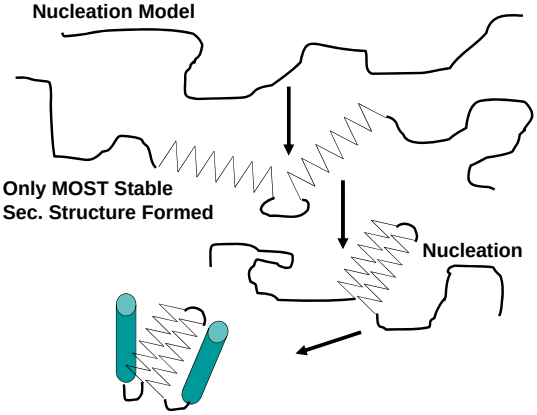
- The "Molten Globule" Model

- "Folding Funnels"

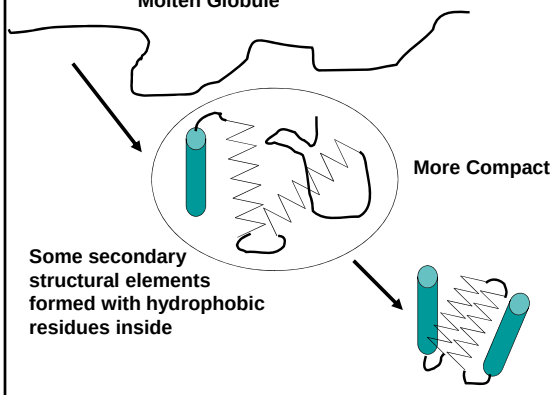
Framework Model



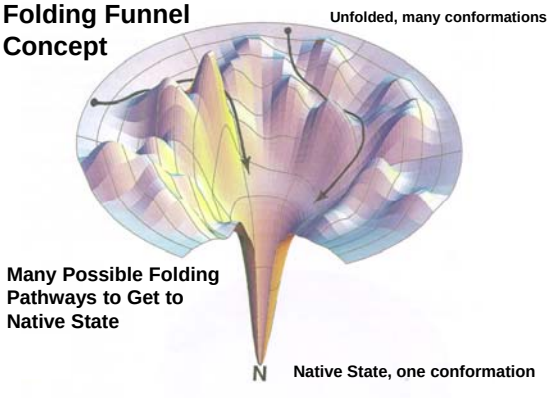
Nucleation Model



Molten Globule

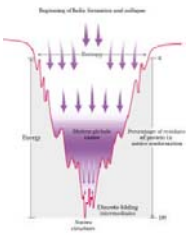


Folding Funnel Concept



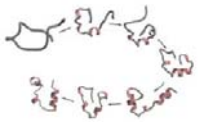
Thermodynamics of Protein Folding

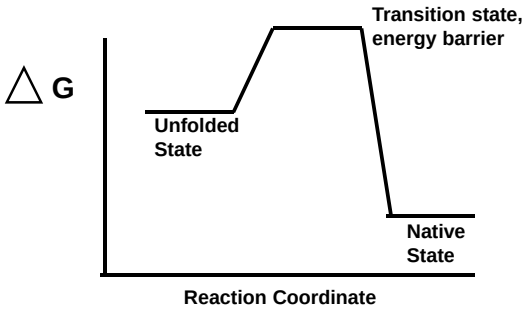
- Bond stretching: 10^{-14} - 10^{-13} sec.
- Elastic vibrations: 10^{-12} - 10^{-11} sec.
- Rotations of surface sidechains: 10^{-11} - 10^{-10} sec.
- Hinge bending: 10^{-11} - 10^{-7} sec.
- Rotation of buried side chains: 10^{-4} - 1 sec.
- Protein folding: 10^{-6} - 10^{2} sec.



Free Energy Funnel

Simulated folding in 1 μsec; peptide in a box of water



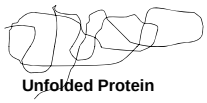


△G, Gibbs Free Energy, the more negative, the more stable the system

Entropy and Enthalpy in Protein Folding

$$\triangle G = \triangle H - T\triangle S$$

bonding flexibility



Unfolded Protein

△H, small, negative
△S, large, positive

Folded Protein

△H, large, negative
△S, small, positive

Compensation in entropy and enthalpy for protein

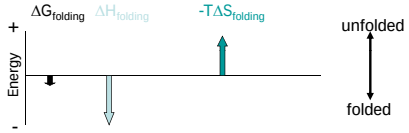
Contribution of entropy of water molecules released upon folding

△S of water is large and positive

Thermodynamics of Protein Folding

$\Delta G_{folding} = G_{folded} - G_{unfolded} =$

$(H_{folded} - H_{unfolded}) - T(S_{folded} - S_{unfolded}) = \Delta H_{folding} - T\Delta S_{folding}$



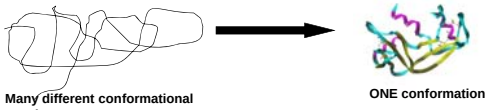
Folded proteins are highly ordered

$\therefore \Delta S_{folding}$ negative, so $-T\Delta S_{folding}$ is a positive quantity

$\Delta H_{folding}$ is a negative quantity - enthalpy is favored in folded state.

Total Gibbs free energy difference is negative – folded state favoured

How Does a Newly Synthesized Protein Go From a Random Coil to the Final Intricately Folded Protein?



Many different conformational species

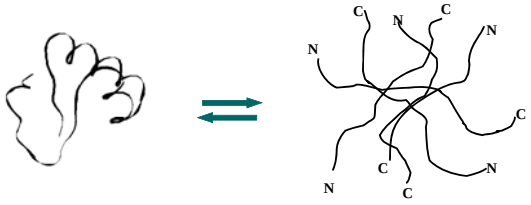
ONE conformation

What are the Forces that Guide this Process?
What are the Steps Involved?
How Fast Can this Happen?

"The native, folded structure of a protein, under optimal conditions, is the most energetically stable conformation possible" Christian Anfinsen, 1972

Native state (N)

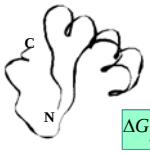
Denatured state (D)



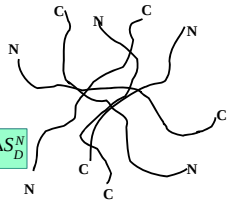
$$\Delta G_D^N = -RT\ln[D]/[N]$$

Stability of the native state is defined as the difference in free energy between the native and denatured states

Native state (N)



Denatured state



$$\Delta G_D^N = \Delta H_D^N - T\Delta S_D^N$$

Size of cavity in solvent ~6500Å²

Average size of cavity in solvent 20,500Å²

ΔS chain: **significantly decreased, due to the well defined conformation**

ΔS chain: **large, due to the large number of different conformations**

Non-bonded interactions: **intra-molecular**

Non-bonded interactions: **inter-molecular**

Compact structure

Non compact structure

Factors that disrupt the Native state

1) **ELECTROLYTE ADDITION**
   - interference with the colloid state

2) **INSOLUBLE SALT FORMATION**
   - Protein+Trichloracetate

3) **ORGANIC SOLVENTS**
   - ETHANOL - interferes with the dielectric constant

4) **HEAT DENATURATION**
   - more energy in system (bonds break)

5) **pH**
   - destroys charge
   - destroys ability to interact with water

6) **DESTRUCTION OF HYDROGEN BONDING**
   - UREA - known H-bond disrupter

Thermodynamic Description of Protein Folding

The native and unfolded states are in equilibrium, the folding reaction can be quantified in terms of thermodynamics.

The equilibrium (N ↔ U) between the native (N) and unfolded (U) states is defined by the equilibrium constant, K, as:

$$K = [U]/[N] = K_U$$

The difference in Gibbs free energy (ΔG) between the unfolded and native states is then:

$$\Delta G = -RT \ln K$$

For K_u, a positive ΔG indicates that the native state is more stable.

Protein Structure Prediction & Alignment

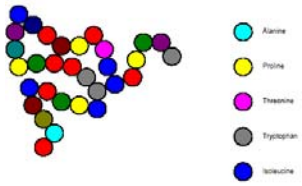
- Protein structure
  - Secondary structure
  - **Tertiary structure**
- Structure prediction
  - Secondary structure
  - **3D structure**
    - **Ab initio**
    - **Comparative modeling**
    - **Threading**
- Structure alignment
  - **3D structure alignment**
  - **Protein docking**

Predicting Protein 3D Structure

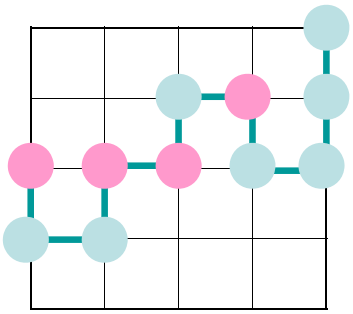
- Goal: Find the best fit of a sequence to a 3D structure
- Ab initio methods
  - Attempt to calculate 3D structure "from scratch"
    - Lattice models
    - off-lattice models
    - Energy minimization
    - Molecular dynamics
- Comparative (homology) modeling
  - Construct 3D model from alignment to protein sequences with known structure
- Threading (fold recognition/reverse folding)
  - Pick best fit to sequences of known 2D/3D structures (folds)

How proteins interact?

- It is believed that *hydrophobic collapse* is a key driving force for protein folding
  - Hydrophobic core!
- Model: A chain of twenty kinds of beads



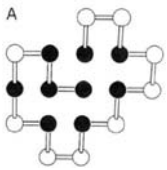
Cubic lattice model



Hydrophobic packing models

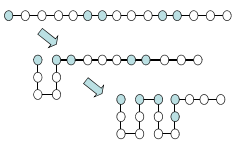
- Dill's HP model
  - Two classes of amino acids, hydrophobic (H) and polar (P)
  - Lattice model for position of amino acids.
  - Thread chain of H's and P's through lattice to maximize number of H-H contacts

2D A



3D B

Hydrophobic Zipper



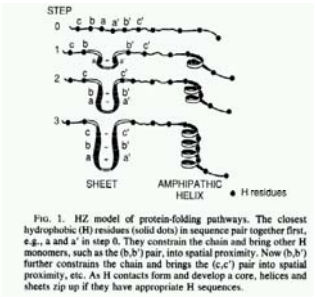


FIG. 1. HZ model of protein-folding pathways. The closest hydrophobic (H) residues (solid dots) in sequence pair together first, e.g., a and a' in step 0. They constrain the chain and bring other H monomers, such as the (b,b') pair, into spatial proximity. Now (b,b') further constrains the chain and brings the (c,c') pair into spatial proximity, etc. As H contacts form and develop a core, helices and sheets zip up if they have appropriate H sequences.

Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 1942–1946, March 1993
Biophysics

Cooperativity in protein-folding kinetics

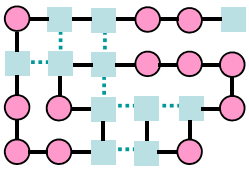
KEN A. DILL, KLAUS M. FIEBIG, AND HUE SUN CHAN

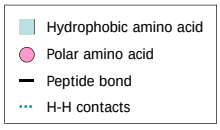
HP Lattice Model

- Simplifications in the model:
  - All amino acids are classified as hydrophobic (H) or polar (P). A protein is represented as a string of H's and P's. HHHHHPPPHHHPP
  - Space is discretized. Each amino acid is embedded to a single lattice point. A protein fold corresponds to a self-avoiding walk over the lattice.
  - The **energy function** is defined as

E = –(# of H-H contacts not including covalent interaction).

Example of HP lattice model





E = -Number of H-H contacts (except for peptide bonds)

= -7

HP Lattice Model

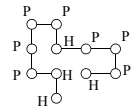
- Other lattices
  - 2D triangular lattice, 3D-diamond lattice
- Other energy functions
  - `HP=0, HH=-1, PP=1`
- Lattice model can be used
  - Study qualitative features of protein folding
  - Reduce search space in structure prediction methods
  - Study potential effectiveness of the methods for structure prediction (inverse folding problem)

Exercise

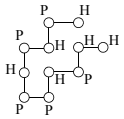
- Find native structures of S_1 and S_2
  - S_1 = HHPPPPHPPPH
  - S_2 = HHPHPPHPHPH
- Thread S_2 on to the structure of S_1 and find the energy associated with that fold

Exercise

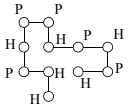
- Find native structures of S_1 and S_2
  - S_1 = HHPPPPHPPPH
  - S_2 = HHPHPPHPHPH
- Thread S_2 on to the structure of S_1 and find the energy associated with that fold



$E(S(S_1), S_1) = -2$;



$E(S(S_2), S_2) = -4$;



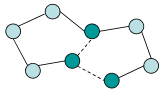
$E(S(S_1), S_2) = -3$.

Summary

- Approach
  - Reduce computation by limiting degrees of freedom
  - Limit α-carbon (Cα) atoms to positions on 2D or 3D lattice
  - Protein sequence → represented as path through lattice points
  - H-P (hydrophobic-polar) cost model
    - Each residue → hydrophobic (H) or hydrophilic (P)
    - Score position of sequence → maximize H-H contacts
- Problem
  - Greatly simplified problem
  - Emphasis on forming
    - hydrophobic core

Off-Lattice Models

- Approach
  - Compromise between lattice model and molecular dynamics
  - Backbone placement → allowed by Ramachandran plot
  - Represent as phi & psi angles of α-carbon atoms
  - Degree of precision
    - α-carbon only
    - All backbone atoms
    - All backbone atoms + side chains (residues)
    - Common conformation (positions) of side chain = rotamer
- Problem
  - Still simplified problem
  - Increased computation cost



Energy Minimization

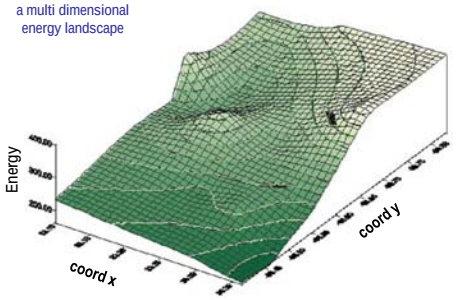
- Hypothesis
  - Amino acids have different chemical/electrical properties
  - Different fold protein have different levels of energy
  - A protein folds into its minimum energy configuration
- Energy function
  - Calculate thermodynamic energy from interatomic forces
    - Hydrophobic contacts, disulfide bond/bridge formation, electrostatic /steric interaction, van der Waals forces, …
- Pseudo-energy function
  - Calculate scoring function based on observed 3D structures
    - Common conformations → low energy
    - Rare/uncommon conformations → very high/high energy

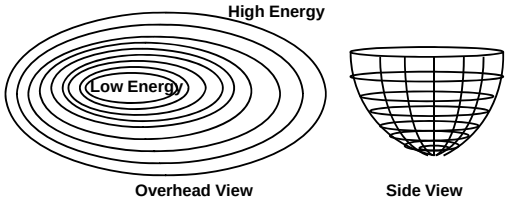
Energy Minimization

- Approach
  - Compute energy of (denatured) protein structure configuration
    - Use energy / pseudo-energy function
  - Incrementally fold protein → reduce energy at each step
    - Model actual observed protein folding process
  - Iterate until convergence to minimum energy
    - Use steepest descent, simulated annealing, etc…
- Problem
  - Energy calculations → expensive
  - Pseudo-energy calculations → heuristics with no physics basis
  - May not be able to converge to correct solution

Potential Energy Surface (PES)

a multi dimensional energy landscape





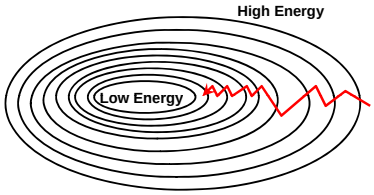
Minimization Methods

- Energy surfaces for proteins are complex hyperdimensional spaces
- Biggest problem is overcoming local minimum problem
- Simple methods (slow) to complex methods (fast)
  - Monte Carlo Method
  - Steepest Descent
  - Conjugate Gradient

Steepest Descent & Conjugate Gradients

- Frequently used for energy minimization of large (and small) molecules
- Ideal for calculating minima for complex (i.e. non-linear) surfaces or functions
- Both use derivatives to calculate the slope and direction of the optimization path
- Both require that the scoring or energy function be differentiable (smooth)

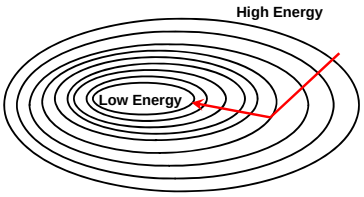
Steepest Descent



Makes small locally steep moves down gradient

The steepest descent method uses the first derivative to determine the direction towards the minimum.

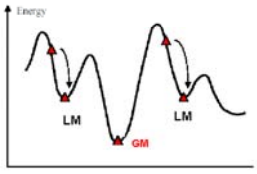
Conjugate Gradient Minimization



Includes information about the prior history of path

Molecular Mechanics - Energy Minimization

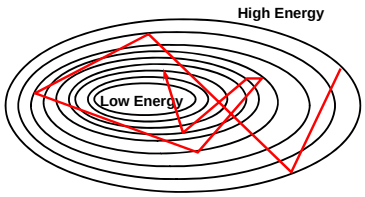
- **The energy of the system is minimized. The system tries to relax**
- **Typically, the system relaxes to a local minimum (LM).**



Monte Carlo Algorithm

- Generate a conformation or alignment (a state)
- Calculate that state's energy or "score"
- If that state's energy is less than the previous state accept that state and go back to step 1
- If that state's energy is greater than the previous state accept it if a randomly chosen number is < $e^{-E/kT}$ where E is the state energy otherwise reject it
- Go back to step 1 and repeat until done

Monte Carlo Minimization



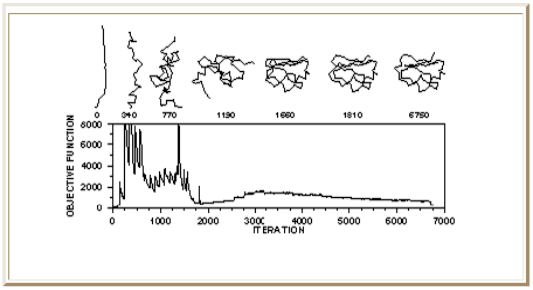
Performs a progressive or directed random search

Energy Minimization

- E = f(x)
- E is a function of coordinates either cartesian or internal
- At minimum the first derivatives are zero and the second derivatives are all positive

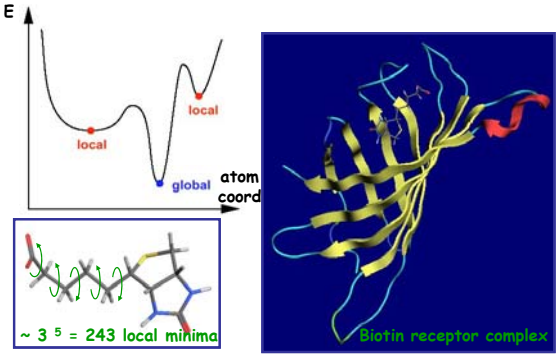
$$\frac{dE}{dx_i} = 0$$

$$\frac{d^2E}{dx_i^2} \rangle 0$$



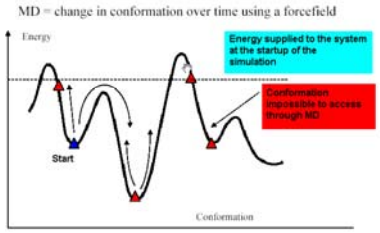
- Treat Protein molecule as a set of balls (with mass) connected by rigid rods and springs
- Rods and springs have empirically determined force constants

Conformation



Molecular Dynamics (MD)

In molecular dynamics, energy is supplied to the system, typically using a constant temperature (i.e. constant average kinetic energy).



Molecular Dynamics (MD)

- Use Newtonian mechanics to calculate the net force and acceleration experienced by each atom.
- Each atom *i* is treated as a point with mass m_i and fixed charge q_i
- Determine the force F_i on each atom:

$$\vec{F}_i = m_i \frac{d^2\vec{r}_i}{dt^2} = -\vec{\nabla} v(\vec{R})$$

- Use positions and accelerations at time *t* (and positions from $t - \delta t$) to calculate new positions at time $t + \delta t$

Initial velocities (v_i)

using the Boltzmann distribution at the given temperature

$v_i = (m_i/2\pi kT)^{1/2} \exp(- m_i v_i^2/2kT)$

Molecular dynamics (MD) simulations

$V_i = \Sigma_k$(energies of interactions between i and all other residues k located within a cutoff distance of R_c from i)

- Derivative of V with respect to the position vector $r_i = (x_i, y_i, z_i)^T$ at each step

$a_{xi} \sim -\partial V/\partial x_i$
$a_{yi} \sim -\partial V/\partial y_i$
$a_{zi} \sim -\partial V/\partial z_i$

Non-Bonded Interaction Potentials

- **Electrostatic interactions of the form $E_{ik}(es) = q_i q_k / r_{ik}$**
- **Van der Waals interactions $E_{ij}(vdW) = - a_{ik}/r_{ik}^6 + b_{ik}/r_{ik}^{12}$**

Bonded Interaction Potentials

- **Bond stretching $E_i(bs) = (k_{bs}/2)(l_i - l_i^0)^2$**
- **Bond angle distortion $E_i(bad) = (k_\theta/2)(\theta_i - \theta_i^0)^2$**
- **Bond torsional rotation $E_i(tor) = (k_\phi/2) f(\cos\phi_i)$**

Implicit Solvent Models

Water molecules are not included as molecules, but represented by an extra potential on the solvent accessible surface.

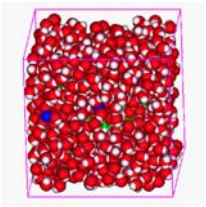
·only 50% slower than vacuum calculations

·~10 times faster than explicit water MD

Explicit Solvent Models

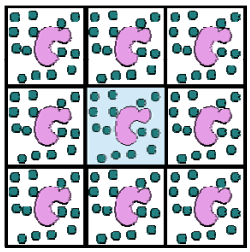
Water molecules are explicitly included as individual molecules.

· Force Fields for water molecules are not trivial …
· Computationally expensive …



Periodic Boundary Conditions (PBC)

· Periodic boundary conditions are used to simulate solvated systems or crystals.
· In solvated systems, PBC prevents that the solvent "evaporates *in silico*"



Molecular dynamics (MD) simulations

Example: gradient of vdW interaction with k, with respect to r_i

- $E_{ik}(vdW) = - a_{ik}/r_{ik}^6 + b_{ik}/r_{ik}^{12}$
- $r_{ik} = r_k - r_i$
  - $x_{ik} = x_k - x_i$
  - $y_{ik} = y_k - y_i$
  - $z_{ik} = z_k - z_i$
  - $r_{ik} = [(x_k - x_i)^2 + (y_k - y_i)^2 + (z_k - z_i)^2]^{1/2}$

$\partial V/\partial x_i = \partial [- a_{ik}/r_{ik}^6 + b_{ik}/r_{ik}^{12}] / \partial x_i$

where $r_{ik}^6 = [(x_k - x_i)^2 + (y_k - y_i)^2 + (z_k - z_i)^2]^3$

Molecular Dynamics

- Goal
  - Provides a way to observe the motion of large molecules such as proteins at the atomic level – dynamic simulation
- Approach
  - Model all interatomic forces acting on atoms in protein
    - Potential energy function (Newtonian mechanics)
  - Perform numerical simulations to predict fold
    - Repeat for each atom at each time step
      - Calculate & add up all (pairwise) forces
        - » bonds:
        - » non-bonded: electrostatic and van der Waals'
      - Apply force, move atom to new position (Newton's 2nd law F = ma)
    - Obtain trajectories of motion of molecule

MD

- Problem with MD
  - Smaller time step → more accurate simulation
  - Modeling folding is computationally intensive
  - Current models require tiny (10^{-15} second) time steps
  - Simulations reported for at most 10^{-6} seconds
  - Folding requires 1 second or more

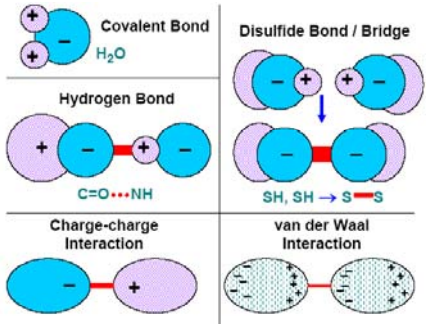
Inter-atomic Forces

- Covalent bond (short range, very strong)
  - Binds atoms into molecules / macromolecules
- Hydrogen bond (short range, strong)
  - Binds two polar groups (hydrogen + electronegative atom)
- Disulfide bond / bridge (short range, very strong)
  - Covalent bond between sulfhydryl (sulfur + hydrogen) groups
  - Sulfhydryl found in cysteine residues
  - Two sulfhydryl groups oxidize → disulfide (S–S) bond
  - Oxidation may require external oxidant (enzyme)
  - Hydrogen & disulfide bonds help stablize 3D protein structure

Inter-atomic Forces

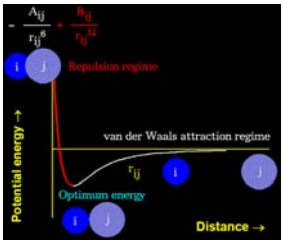
- Hydrophobic/hydrophillic interaction (weak)
  - Hydrogen bonding with H2O in solution
    - Non-polar residues interfere (hydrophobic)
    - Polar residues participate (hydrophillic)
  - Main cause of globular 3D protein → protect hydrophobic core
- Charge-charge, charge-dipole, dipole-dipole (weak)
  - Electrostatic attractive force
- Van der Waal's interaction (very weak)
  - Nonspecific electrostatic attractive force
  - From transitive attractions between instantaneous dipoles
- Steric interaction (very short range, very strong)
  - Repulsive force between atomic nuclei

Types of Inter-atomic Forces



Lennard-Jones Potential

- Forces
  - Van der Waal's (attractive, far)
  - Steric interaction (repulsive, close)
- Lennard-Jones
  - Plot of pair potential energy vs. distance
  - Local minima (energy well) is stable distance for two atoms



Potential Energy

- Components

  (1) bond length

  Bonds behave like spring with equilibrium bond length depending on bond type. Increase or decrease from equilibrium length requires higher energy.

$$E_{pot} = \sum_{b} K_2 (b - b_0)^2 \quad (1)$$

Potential Energy

(2) bond angle

- Bond angles have equilibrium value eg 108 for H-C-H
- Behave as if sprung.

$$E_{pot} = \sum_{\theta} H_{\theta}(\theta - \theta_0)^2 \quad (2)$$

- Increase or decrease in angle requires higher energy.

Potential Energy

(3) torsion angle

Rotation can occur about single bond in A-B-C-D but energy depends on torsion angle (angle between CD & AB viewed along BC). Staggered conformations (angle +60, -60 or 180 are preferred).

$$E_{pot} = \sum_{\phi} \frac{V_n}{2}[1 + \cos(n\phi - \phi_0)] \quad (3)$$

Potential Energy

(4) van der Waals interactions

Interactions between atoms not near neighbours expressed by Lennard-Jones potential. Very high repulsive force if atoms closer than sum of van der Waals radii. Attractive force if distance greater. Because of strong distance dependence, van der Waals interactions become negligible at distances over 15 Å.

$$E_{pot} = \sum \varepsilon[(r^*/r)^{12} - 2(r^*/r)^6] \quad (4)$$

Potential Energy

(5) Electrostatic interactions

All atoms have partial charge eg in C=O, C has partial positive charge, O atom partial negative charge. Two atoms that have the same charge repel one another, those with unlike charge attract.

$$E_{pot} = \sum q_i q_j / \varepsilon_{ij} r_{ij} \quad (5)$$

Electrostatic energy falls off much less quickly than for van der Waals interactions and may not be negligible even at 30 Å.

Potential Energy

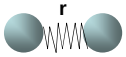
- Potential Energy is given by the sum of these contributions:

$$E_{pot} = \underbrace{\sum_{b} K_2(b - b_0)^2}_{(1)} + \underbrace{\sum_{\theta} H_{\theta}(\theta - \theta_0)^2}_{(2)} + \underbrace{\sum_{\phi} \frac{V_n}{2}[1 + \cos(n\phi - \phi_0)]}_{(3)} + \underbrace{\sum \varepsilon[(r^*/r)^{12} - 2(r^*/r)^6]}_{(4)} + \underbrace{\sum q_i q_j / \varepsilon_{ij} r_{ij}}_{(5)} + \underbrace{\sum \left[\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right]}_{(6)}$$

- Hydrogen bonds are usually supposed to arise by electrostatic interactions but occasionally a small extra term is added.

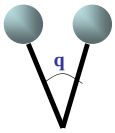
Energy Terms

Covalent



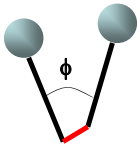
Stretching

$K_r(r_i - r_j)^2$



Bending

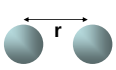
$K_\theta(\theta_i - \theta_j)^2$



Torsional

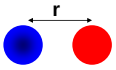
$K_\phi(1-\cos(n\phi_j))^2$

Noncovalent



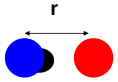
van der Waals

$A_{ij}/r^6 - B_{ij}/r^{12}$



Coulomb

$q_i q_j / 4\pi\varepsilon r_{ij}$



H-bond

$C_{ij}/r^{10} - D_{ij}/r^{12}$

Interaction	*Approx. bond strength in kJ/mole*
Covalent bonds	**> 200 (ranging up to 900)**
Ionic	**20-40**
Hydrogen bond	**~5-10**
Hydrophobic	**~ 8**
van der Waals	**~ 4**

AMBER (*Assisted Model Building with Energy Refinement*) force field

$$E_{total} = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \sum_{1}^{3} \frac{V_n}{2}[1 + \cos(n\omega)]$$
$$+ \sum_{i<j}^{atoms} \left(\frac{a_{ij}}{r_{ij}^{12}} - \frac{b_{ij}}{r_{ij}^{6}} \right) + \sum_{i<j}^{atoms} \frac{q_i q_j}{\varepsilon r_{ij}}$$

Force fields

- A force field is the description of how potential energy depends on parameters
- Several force fields are available
  - AMBER used for proteins and nucleic acids (UCSF)
  - CHARMM (Harvard)
  - …
- Force fields differ:
  - in the precise form of the equations
  - in values of the constants for each atom type

Force Field Parameterization

- Equilibrium bond distances and angles: X-ray crystallography
- Bond and angle force constants: vibrational spectra, normal mode calculations with QM
- Dihedral angle parameters: difficult to measure directly experimentally; fit to QM calculations for rotations around a bond with other motions fixed
- Atom charges: fit to experimental liquid properties, ESP charge fitting to reproduce electrostatic potentials of high level QM, X-ray crystallographic electron density
- Lennard-Jones parameters: often most difficult to determine, fit to experimental liquid properties, intermolecular energy fitting

Applications

- NMR or X-ray structure refinement
- Protein structure prediction
- Protein folding kinetics and mechanics
- Conformational dynamics
- Global optimization
- DNA/RNA simulations
- Membrane proteins/lipid layers simulations

Which Force Field to Use?

- Most popular force fields: CHARMM, AMBER and OPLSAA
- OPLSAA(2000): Probably the best available force field for condensed-phase simulation of peptides. Work to develop parameterization that will include broader classes of drug-like molecules is ongoing. GB/SA solvation energies are good.
- MMFF: An excellent force field for biopolymers and many drug-like organic molecules that do not have parameters in other force fields.
- AMBER*/OPLS*: Good force fields for biopolymers and carbohydrates; many parameters were added in MacroModel which extend the scope of this force field to a number of important organic functional groups. GB/SA solvation energies range from moderate (AMBER*) to good (OPLS*).
- AMBER94: An excellent force field for proteins and nucleic acids. However, there are no extensions for non-standard residues or organic molecules, also there is a alpha-helix tendency for proteins. AMBER99 fixes this helix problem to some degree, but not completely.
- MM2*/MM3*: Excellent force fields for hydrocarbons and molecules with single or remotely spaced functional groups. GB/SA solvation energies tend to be poor relative to those calculated with other force fields.
- CHARMM22: Good general purpose force field for proteins and nucleic acids. A bit weak for drug-like organic molecules.
- GROMOS96: Good general purpose force field for proteins, particularly good for free energy perturbations due to soft-core potentials. Weak for reproducing solvation free energies of organic molecules and small peptides.

More on Potential

- To reduce the complexity of calculations atoms grouped into types (potential atom types)
  - all H´s in methane are the same & similar to H´s in ethane
  - the C atoms in ethane are different from those in ethylene
  - the O in a C=O group is different from the O in a C-O-H group. But O atoms in alcohols are similar.

Atom types (AMBER)

Table 1. List of Atom Types[a]

atom	type	description
carbon	CT	any sp^3 carbon
	C	any carbonyl sp^2 carbon
	CA	any aromatic sp^2 carbon and (Cε of Arg)
	CM	any sp^2 carbon, double bonded
	CC	sp^2 aromatic in 5-membered ring with one substituent + next to nitrogen (Cγ in His)
	CV	sp^2 aromatic in 5-membered ring next to carbon and lone pair nitrogen (e.g. Cδ in His (δ))
	CW	sp^2 aromatic in 5-membered ring next to carbon and NH (e.g. Cδ in His (ε) and in Trp)
	CR	sp^2 aromatic in 5-membered ring next to two nitrogens (Cγ and Cε in His)
	CB	sp^2 aromatic at junction of 5- and 6-membered rings (Cδ in Trp) and both junction atoms in Ade and Gua
	C*	sp^2 aromatic in 5-membered ring next to two carbons (e.g. Cγ in Trp)
	CN	sp^2 junction between 5- and 6-membered rings and bonded to CH and NH (Cε in Trp)
	CK	sp^2 carbon in 5-membered aromatic between N and N-R (C8 in purines)
	CQ	sp^2 carbon in 6-membered ring between lone pair nitrogens (e.g. C2 in purines)
nitrogen	N	sp^2 nitrogen in amides
	NA	sp^2 nitrogen in aromatic rings with hydrogen attached (e.g. protonated His, Gua, Trp)
	NB	sp^2 nitrogen in 5-membered ring with lone pair (e.g. N7 in purines)
	NC	sp^2 nitrogen in 6-membered ring with lone pair (e.g. N3 in purines)
	N*	sp^2 nitrogen in 5-membered ring with carbon substituent (in purine nucleosides)
	N2	sp^2 nitrogen of aromatic amines and guanidinium ions
	N3	sp^3 nitrogen
oxygen	OW	sp^3 oxygen in TIP3P water
	OH	sp^3 oxygen in alcohols, tyrosine, and protonated carboxylic acids
	OS	sp^3 oxygen in ethers
	O	sp^2 oxygen in amides
	O2	sp^2 oxygen in anionic acids
sulfur	S	sulfur in methionine and cysteine
	SH	sulfur in cysteine
phosphorus	P	phosphorus in phosphates
hydrogen	H	H attached to N
	HW	H in TIP3P water
	HO	H in alcohols and acids
	HS	H attached to sulfur
	HA	H attached to aromatic carbon
	HC	H attached to aliphatic carbon with no electron-withdrawing substituents
	H1	H attached to aliphatic carbon with one electron-withdrawing substituent
	H2	H attached to aliphatic carbon with two electron-withdrawing substituents
	H3	H attached to aliphatic carbon with three electron-withdrawing substituents
	HP	H attached to carbon directly bonded to formally positive atoms (e.g. C next to NH_3^+ of lysine)
	H4	H attached to aromatic carbon with one electronegative neighbor (e.g. hydrogen on C5 of Trp, C6 of Thy)
	H5	H attached to aromatic carbon with two electronegative neighbors (e.g. H8 of Ade and Gua and H2 of Ade)

Bond Parameters

Bond Parameters

bond	K_r	r_{eq}	bond	K_r	r_{eq}	bond	K_r	r_{eq}	bond	K_r	r_{eq}
C-CA	469.0	1.409	CA-HA	367.0	1.080	CM-HA	367.0	1.080	CT-S	227.0	1.810
C-CB	447.0	1.419	CA-N2	481.0	1.340	CM-N*	448.0	1.365	CT-SH	237.0	1.810
C-CM	410.0	1.444	CA-NA	427.0	1.381	CN-NA	428.0	1.380	CV-H4	367.0	1.080
C-CT	317.0	1.522	CA-NC	483.0	1.339	CQ-H5	367.0	1.080	CV-NB	410.0	1.394
C-N	490.0	1.335	CB-CB	520.0	1.370	CQ-NC	502.0	1.324	CW-H4	367.0	1.080
C-N*	424.0	1.383	CB-CN	447.0	1.419	CR-H5	367.0	1.080	CW-NA	427.0	1.381
C-NA	418.0	1.388	CB-N*	436.0	1.374	CR-NA	477.0	1.343	H-N	434.0	1.010
C-NC	457.0	1.358	CB-NB	414.0	1.391	CR-NB	488.0	1.335	H-N*	434.0	1.010
C-O	570.0	1.229	CB-NC	461.0	1.354	CT-CT	310.0	1.526	H-N2	434.0	1.010
C-O2	656.0	1.250	CC-CT	317.0	1.504	CT-F	367.0	1.380	H-N3	434.0	1.010
C-OH	450.0	1.364	CC-CV	512.0	1.375	CT-H1	340.0	1.090	H-NA	434.0	1.010
C*-CB	388.0	1.459	CC-CW	518.0	1.371	CT-H2	340.0	1.090	HO-OH	553.0	0.960
C*-CT	317.0	1.495	CC-NA	422.0	1.385	CT-H3	340.0	1.090	HO-OS	553.0	0.960
C*-CW	546.0	1.352	CC-NB	410.0	1.394	CT-HC	340.0	1.090	HS-SH	274.0	1.336
C*-HC	367.0	1.080	CK-H5	367.0	1.080	CT-HP	340.0	1.090	O2-P	525.0	1.480
CA-CA	469.0	1.400	CK-N*	440.0	1.371	CT-N	337.0	1.449	OH-P	230.0	1.610
CA-CB	469.0	1.404	CK-NB	529.0	1.304	CT-N*	337.0	1.475	OS-P	230.0	1.610
CA-CM	427.0	1.433	CM-CM	549.0	1.350	CT-N2	337.0	1.463	OW-HW	553.0	0.9572
CA-CN	469.0	1.400	CM-CT	317.0	1.510	CT-N3	367.0	1.471	S-S	166.0	2.038
CA-CT	317.0	1.510	CM-H4	367.0	1.080	CT-OH	320.0	1.410			
CA-H4	367.0	1.080	CM-H5	367.0	1.080	CT-OS	320.0	1.410			

Angle Parameters

Angle Parameters

angle	K_θ	θ_{eq}	angle	K_θ	θ_{eq}	angle	K_θ	θ_{eq}	angle	K_θ	θ_{eq}
C-CA-CA	63.0	120.00	CA-CT-HC	50.0	109.50	CN-NA-H	30.0	123.10	H1-CT-N2	50.0	109.50
C-CA-HA	35.0	120.00	CA-N2-CT	50.0	123.20	CR-NA-CW	70.0	120.00	H1-CT-OH	50.0	109.50
C-CB-CB	63.0	119.20	CA-N2-H	35.0	120.00	CR-NA-H	30.0	120.00	H1-CT-OS	50.0	109.50
C-CB-NB	70.0	130.00	CA-NA-H	30.0	118.00	CR-NB-CV	70.0	117.00	H1-CT-S	50.0	109.50
C-CM-CM	63.0	120.70	CA-NC-CB	70.0	112.20	CT-C-N	70.0	116.60	H1-CT-SH	50.0	109.50
C-CM-CT	70.0	119.70	CA-NC-CQ	70.0	118.60	CT-C-O	80.0	120.40	H2-CT-H2	35.0	109.50
C-CM-H4	35.0	119.70	CB-C-NA	70.0	111.30	CT-C-O2	70.0	117.00	H2-CT-N*	50.0	109.50
C-CM-HA	35.0	119.70	CB-C-O	80.0	128.80	CT-C*-CW	70.0	125.00	H2-CT-OS	50.0	109.50
C-CT-CT	63.0	111.10	CB-C*-CT	70.0	128.60	CT-CC-CV	70.0	120.00	H4-CM-N*	35.0	119.10
C-CT-H1	50.0	109.50	CB-C*-CW	63.0	106.40	CT-CC-CW	70.0	120.00	H4-CV-NB	35.0	120.00
C-CT-HC	50.0	109.50	CB-CA-H4	35.0	120.00	CT-CC-NA	70.0	120.00	H4-CW-NA	35.0	120.00
C-CT-HP	50.0	109.50	CB-CA-HA	35.0	120.00	CT-CC-NB	70.0	120.00	H5-CK-N*	35.0	123.05
C-CT-N	63.0	110.10	CB-CA-N2	70.0	123.50	CT-CT-CT	40.0	109.50	H5-CK-NB	35.0	123.05
C-CT-N3	80.0	111.20	CB-CA-NC	70.0	117.30	CT-CT-H1	50.0	109.50	H5-CQ-NC	35.0	115.45
C-N-CT	50.0	121.90	CB-CB-N*	70.0	106.20	CT-CT-H2	50.0	109.50	H5-CR-NA	35.0	120.00
C-N-H	30.0	120.00	CB-CB-NB	70.0	110.40	CT-CT-HC	50.0	109.50	H5-CR-NB	35.0	120.00
C-N*-CM	70.0	121.60	CB-CB-NC	70.0	127.70	CT-CT-HP	50.0	109.50	HC-CT-HC	35.0	109.50
C-N*-CT	70.0	117.60	CB-CN-NA	70.0	104.40	CT-CT-N	80.0	109.70	HO-OH-P	45.0	108.50
C-N*-H	30.0	119.20	CB-N*-CK	70.0	105.40	CT-CT-N*	50.0	109.50	HP-CT-HP	35.0	109.50
C-NA-C	70.0	126.40	CB-N*-CT	70.0	125.80	CT-CT-N2	80.0	111.20	HP-CT-N3	50.0	109.50
C-NA-CA	70.0	125.20	CB-N*-H	30.0	125.80	CT-CT-N3	80.0	111.20	HS-SH-HS	35.0	92.07
C-NA-H	30.0	116.80	CB-NB-CK	70.0	103.80	CT-CT-OH	50.0	109.50	HW-OW-HW	100.0	104.52
C-NC-CA	70.0	120.50	CB-NC-CQ	70.0	111.00	CT-CT-OS	50.0	109.50	N-C-O	80.0	122.90
C-OH-HO	35.0	113.00	CC-CT-CT	63.0	113.10	CT-CT-S	50.0	114.70	N*-C-NA	70.0	115.40
C*-CB-CA	63.0	134.90	CC-CT-HC	50.0	109.50	CT-CT-SH	50.0	108.60	N*-C-NC	70.0	118.60
C*-CB-CN	63.0	108.80	CC-CV-H4	35.0	120.00	CT-N-CT	50.0	118.00	N*-C-O	80.0	120.90
C*-CT-CT	63.0	115.60	CC-CV-NB	70.0	120.00	CT-N-H	30.0	118.04	N*-CB-NC	70.0	126.20
C*-CT-HC	50.0	109.50	CC-CW-H4	35.0	120.00	CT-N2-H	35.0	118.40	N*-CK-NB	70.0	113.90
C*-CW-H4	35.0	120.00	CC-CW-NA	70.0	120.00	CT-N3-H	50.0	109.50	N*-CT-OS	50.0	109.50
C*-CW-NA	70.0	108.70	CC-NA-CR	70.0	120.00	CT-OH-HO	55.0	108.50	N2-CA-N2	70.0	120.00
CA-C-CA	63.0	120.00	CC-NA-H	30.0	120.00	CT-OS-CT	60.0	109.50	N2-CA-NA	70.0	116.00
CA-C-OH	70.0	120.00	CC-NB-CR	70.0	117.00	CT-OS-P	100.0	120.50	N2-CA-NC	70.0	119.30
CA-CA-CA	63.0	120.00	CK-N*-CT	70.0	128.80	CT-S-CT	62.0	98.90	NA-C-O	80.0	120.60
CA-CA-CB	63.0	120.00	CK-N*-H	30.0	128.80	CT-S-S	68.0	103.70	NA-CA-NC	70.0	123.30
CA-CA-CN	63.0	120.00	CM-C-NA	70.0	114.10	CT-SH-HS	43.0	96.00	NA-CR-NA	70.0	120.00
CA-CA-CT	70.0	120.00	CM-C-O	80.0	125.30	CV-CC-NA	70.0	120.00	NA-CR-NB	70.0	120.00
CA-CA-H4	35.0	120.00	CM-CA-N2	70.0	120.10	CW-CC-NA	70.0	120.00	NC-C-O	80.0	122.50

Torsion Parameters

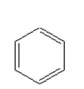
Torsional Parameters

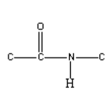
torsion	no. of paths	$V_n/2$	γ	n	torsion	no. of paths	$V_n/2$	γ	n
X-C-CA-X	4	14.50	180.0	2.0	X-CT-OH-X	3	0.50	0.0	3.0
X-C-CB-X	4	12.00	180.0	2.0	X-CT-OS-X	3	1.15	0.0	3.0
X-C-CM-X	4	8.70	180.0	2.0	X-CT-S-X	3	1.00	0.0	3.0
X-C-CT-X	4	0.00	0.0	2.0	X-CT-SH-X	3	0.75	0.0	3.0
X-C-N-X	4	10.00	180.0	2.0	X-CV-NB-X	2	4.80	180.0	2.0
X-C-N*-X	4	5.80	180.0	2.0	X-CW-NA-X	4	6.00	180.0	2.0
X-C-NA-X	4	5.40	180.0	2.0	X-OH-P-X	3	0.75	0.0	3.0
X-C-NC-X	2	8.00	180.0	2.0	X-OS-P-X	3	0.75	0.0	3.0
X-C-OH-X	2	1.80	180.0	2.0	C-N-CT-C	1	0.0	0.0	−4.0
X-C*-CB-X	4	6.70	180.0	2.0	C-N-CT-C	1	0.0	180.0	−3.0
X-C*-CT-X	6	0.00	0.0	2.0	C-N-CT-C	1	0.20	180.0	−2.0
X-C*-CW-X	4	26.10	180.0	2.0	C-N-CT-C	1	0.00	180.0	1.0
X-CA-CA-X	4	14.50	180.0	2.0	CT-CT-C-N	1	0.100	0.0	−4.0
X-CA-CB-X	4	14.00	180.0	2.0	CT-CT-C-N	1	0.000	0.0	−3.0
X-CA-CM-X	4	10.20	180.0	2.0	CT-CT-C-N	1	0.07	0.0	−2.0
X-CA-CN-X	4	14.50	180.0	2.0	CT-CT-C-N	1	0.000	180.0	1.0
X-CA-CT-X	6	0.00	0.0	2.0	CT-CT-N-C	1	0.50	180.0	−4.0
X-CA-N2-X	4	9.60	180.0	2.0	CT-CT-N-C	1	0.15	180.0	−3.0
X-CA-NA-X	4	6.00	180.0	2.0	CT-CT-N-C	1	0.00	180.0	−2.0
X-CA-NC-X	2	9.60	180.0	2.0	CT-CT-N-C	1	0.53	0.0	1.0
X-CB-CB-X	4	21.80	180.0	2.0	CT-CT-OS-CT	1	0.383	0.0	−3.0
X-CB-CN-X	4	12.00	180.0	2.0	CT-CT-OS-CT	1	0.1	180.0	2.0
X-CB-N*-X	4	6.60	180.0	2.0	CT-S-S-CT	1	0.60	0.0	3.0
X-CB-NB-X	2	5.10	180.0	2.0	CT-S-S-CT	1	3.50	0.0	−2.0
X-CB-NC-X	2	8.30	180.0	2.0	H-N-C-O	1	2.50	180.0	−2.0
X-CC-CT-X	6	0.00	0.0	2.0	H-N-C-O	1	2.00	0.0	1.0
X-CC-CV-X	4	20.60	180.0	2.0	N-CT-C-N	1	0.40	180.0	−4.0
X-CC-CW-X	4	21.50	180.0	2.0	N-CT-C-N	1	0.0	0.0	−3.0
X-CC-NA-X	4	5.60	180.0	2.0	N-CT-C-N	1	1.35	180.0	−2.0
X-CC-NB-X	2	4.80	180.0	2.0	N-CT-C-N	1	0.75	180.0	1.0
X-CK-N*-X	4	6.80	180.0	2.0	OH-CT-CT-OH	1	0.144	0.0	−3.0
X-CK-NB-X	2	20.00	180.0	2.0	OH-CT-CT-OH	1	1.00	0.0	2.0
X-CM-CM-X	4	26.60	180.0	2.0	OH-P-OS-CT	1	0.25	0.0	−3.0
X-CM-CT-X	6	0.00	0.0	3.0	OH-P-OS-CT	1	1.20	0.0	2.0
X-CM-N*-X	4	7.40	180.0	2.0	OS-CT-CT-OH	1	0.144	0.0	−3.0
X-CN-NA-X	4	6.10	180.0	2.0	OS-CT-CT-OH	1	1.00	0.0	2.0
X-CQ-NC-X	2	13.60	180.0	2.0	OS-CT-CT-OS	1	0.144	0.0	−3.0

Improper Torsions

Improper Torsions

torsion	$V_n/2$	γ	n	torsion	$V_n/2$	γ	n	torsion	$V_n/2$	γ	n
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]





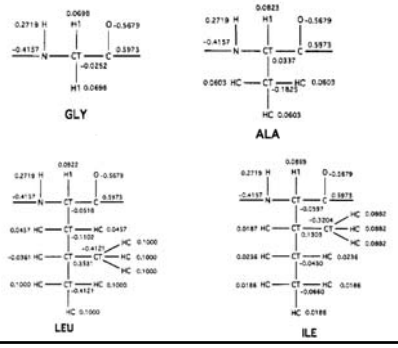
Van der Waals (LJ) Parameters

Van der Waals Parameters

atom type	R^*	ϵ	atom type	R^*	ϵ	atom type	R^*	ϵ	atom type	R^*	ϵ
C	1.9080	0.0860	H2	1.2870	0.0157	HS	0.6000	0.0157	O2	1.6612	0.2100
CA	1.9080	0.0860	H3	1.1870	0.0157	HW	0.0000	0.0000	OH	1.7210	0.2104
CM	1.9080	0.0860	H4	1.4090	0.0150	IP	1.8680	0.00277	OS	1.6837	0.1700
Cs	3.3950	0.0000806	H5	1.3590	0.0150	K	2.6580	0.000328	OW	1.7683	0.1520
CT	1.9080	0.1094	HA	1.4590	0.0150	Li	1.1370	0.0183	P	2.1000	0.2000
F	1.75	0.061	HC	1.4870	0.0157	N*	1.8240	0.1700	Rb	2.9560	0.00017
H	0.6000	0.0157	HO	0.0000	0.0000	N3	1.875	0.1700	S	2.0000	0.2500
H1	1.3870	0.0157	HP	1.1000	0.0157	O	1.6612	0.2100	SH	2.0000	0.2500

$$V_{LJ} = \sum_{i<j} 4\varepsilon_{ij} \left[\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right]$$

Atomic Partial Charges



Protein Structure Prediction and Protein Folding

Protein Structure Prediction

Fundamental Questions

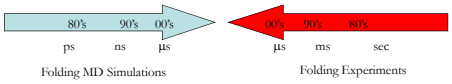
- What is the structure of this protein?
  - Can be experimentally determined, today we know the structure of ~35,000 proteins
  - Can be predicted for some proteins, usually in ~1 day on today's computers

Protein Folding

- How does this protein form this structure?
  - The process or mechanism of folding
  - Limited experimental characterization
- Why does this protein form ***this*** structure?
  - Why not some other fold?
  - Why so quickly? -> Levinthal's Paradox: As there are an astronomical number of conformations possible, an unbiased search would take too long for a protein to fold. Yet most proteins fold in less than a second!

Protein Folding: Fast Folders

Time Scale:

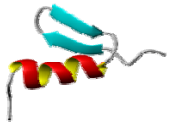


- Trp-cage, designed mini-protein (20 aa): 4µs
- β-hairpin of C-terminus of protein G (16 aa) : 6µs
- Engrailed homeodomain (En-HD) (61 aa): ~27µs
- WW domains (38-44 aa): >24µs
- Fe(II) cytochrome b_{562} (106 aa): extrapolated ~5µs
- B domain of protein A (58 aa): extrapolated ~8µs

Structure Prediction Methods

```
1  QQYTA KIKGR
11 TFRNE KELRD
21 FIEKF KGR
```

Algorithm →



- **Secondary structure (only sequence)**
- **Homology modeling (using related structure)**
- **Fold recognition**
- ***Ab-initio* 3D prediction**

Homology Modeling

- **Assumes similar (homologous) sequences have very similar tertiary structures**
- **Basic structural framework is often the same (same secondary structure elements packed in the same way)**
- **Loop regions differ**
  - **Wide differences possible, even among closely related proteins**

Threading

- Given:
  - sequence of protein P with unknown structure
  - Database of known folds
- Find:
  - Most plausible fold for P
  - Evaluate quality of such arrangement
- Places the residues of unknown P along the backbone of a known structure and determine stability of side chains in that arrangement

Strategies for Protein Structure Prediction

	Comparative Modeling	Fold Recognition	Ab Initio
Method	1. Identify sequence homologs as templates 2. Use sequence alignment to generate model 3. Fill in unaligned regions 4. Improves with data	1. Fold classification 2 3D-Profiles 3. Improves with data	1. Representation 2. Force field 3. Global Optimization 4. Structure at global minimum 5. Can discover new folds
Drawbacks	1. Requires > 25% sequence identity 2. Loops and sidechain conformations are critical	1. Needs good number of proteins in each fold 2. Critically dependent on scoring function	1. Computationally intensive 2. Physical modeling
Resolution	< 3 A	3 - 7 A	> 5 A
Time to Compute	< Day	~ Day	>> Day

Complementarity of the Methods

- **X-ray crystallography- highest resolution structures; faster than NMR**
- **NMR- enables widely varying solution conditions; characterization of motions and dynamic, weakly interacting systems**
- **Computation- fundamental understanding of structure, dynamics and interactions; models without experiment; very fast**

Molecular Dynamics

- Molecular dynamics simulation uses the force field to create a movie of the protein changing with time. With the trajectories obtained, one can:
  - Simulate motions and view the size and time scale of the motions and their correlations
  - Obtain equilibrium properties of the system with appropriate ensemble average
  - Find the global optimum structure using simulated annealing
  - Chart the temperature (salt concentration, …) dependence of the system
  - …

Obtain Trajectory

- Start with a initial structure (Ex. Structure from PDB)
- Assign random starting velocities to the atoms
- Calculating the forces acting on each atom
  - Bonds, non-bonded (electrostatic and van der Val's)
- Numerically integrate **Newton's equations of motion**
  - Verlet method
  - Leapfrog method

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}(t+\tfrac{1}{2}\delta t)\delta t$$

$$\mathbf{v}(t+\tfrac{1}{2}\delta t) = \mathbf{v}(t-\tfrac{1}{2}\delta t) + \tfrac{1}{m}\mathbf{F}(t)\delta t$$

- After equilibrating the system, record the positions and momentum of the atoms as a function of time

Molecular Dynamics

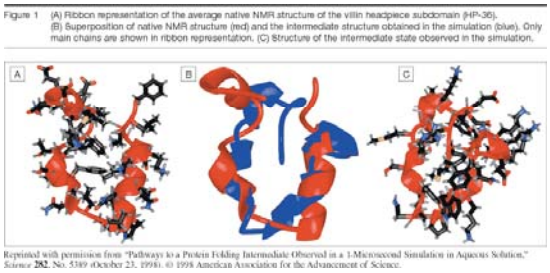
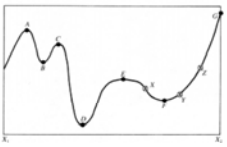


Figure 1 (A) Ribbon representation of the average native NMR structure of the villin headpiece subdomain (HP-36). (B) Superposition of native NMR structure (red) and the intermediate structure obtained in the simulation (blue). Only main chains are shown in ribbon representation. (C) Structure of the intermediate state observed in the simulation.

Reprinted with permission from "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution," *Science* **282**, No. 5389 (October 23, 1998). © 1998 American Association for the Advancement of Science.

Molecular Dynamics

- Energy minimization gives local minimum, not necessarily global minimum.



- Give molecule thermal energy so can explore conformational space & overcome energy barriers.
- Give atoms initial velocity random value + direction. Scale velocities so total kinetic energy =3/2kT * number atoms
- Solve equation of motion to work out position of atoms at 1 fs.

Implicit Solvent Models

Water molecules are not included as molecules, but represented by an extra potential on the solvent accessible surface.

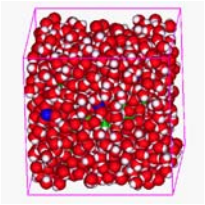
·only 50% slower than vacuum calculations

·~10 times faster than explicit water MD

Explicit Solvent Models

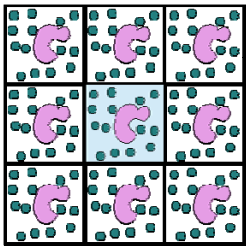
Water molecules are explicitly included as individual molecules.

· Force Fields for water molecules are not trivial …
· Computationally expensive …



Periodic Boundary Conditions (PBC)

· Periodic boundary conditions are used to simulate solvated systems or crystals.
· In solvated systems, PBC prevents that the solvent "evaporates *in silico*"



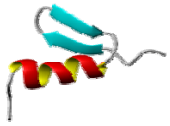
Protein Structure Prediction: *Why Attempt It?*

- A good guess is better than nothing!
  - Enables the design of experiments
  - Potential for high-throughput
- Crystallography and NMR don't always work!
  - Many important proteins do not crystallize
  - Size limitations with NMR

Structure Prediction Methods

```
1  QQYTA KIKGR
11 TFRNE KELRD
21 FIEKF KGR
```

Algorithm →



- Secondary structure (only sequence)
- Homology modeling (using related structure)
- Fold recognition
- *Ab-initio* 3D prediction

Homology Modeling

- Assumes similar (homologous) sequences have very similar tertiary structures
- Basic structural framework is often the same (same secondary structure elements packed in the same way)
- Loop regions differ
  - Wide differences possible, even among closely related proteins

Complementarity of the Methods

- X-ray crystallography- highest resolution structures; faster than NMR
- NMR- enables widely varying solution conditions; characterization of motions and dynamic, weakly interacting systems
- Computation- fundamental understanding of structure, dynamics and interactions; models without experiment

Typical Time Scales ....

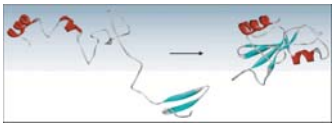
- Bond stretching: 10^{-14} - 10^{-13} sec.
- Elastic vibrations: 10^{-12} - 10^{-11} sec.
- Rotations of surface sidechains: 10^{-11} - 10^{-10} sec.
- Hinge bending: 10^{-11} - 10^{-7} sec.
- Rotation of buried side chains: 10^{-4} - 1 sec.
- Protein folding: 10^{-6} - 10^{2} sec.

Timescale in MD:

- A Typical timestep in MD is 1 fs (10^{-15} sec)

(ideally 1/10 of the highest frequency vibration)

***Ab initio* protein folding simulation**



Physical time for simulation	10^{-4} seconds
Typical time-step size	10^{-15} seconds
Number of MD time steps	10^{11}
Atoms in a typical protein and water simulation	32,000
Approximate number of interactions in force calculation	10^{9}
Machine instructions per force calculation	1000
Total number of machine instructions	10^{23}

BlueGene capacity (floating point operations per second) 1 pentaflop (10^{15})