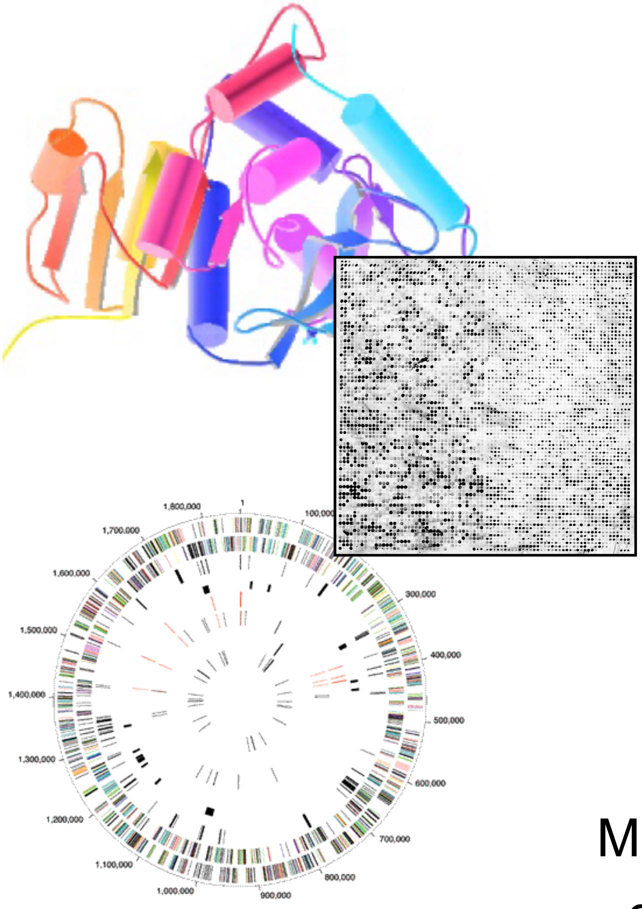
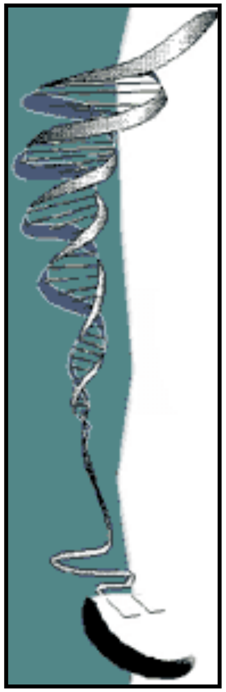
BIOINFORMATICS
Simulation





Mark Gerstein, Yale University
gersteinlab.org/courses/452
(last edit in fall '06, handout version, including in-class changes)

Start of class M8 [2006,11.27]

Overview:
Electrostatics + Basic Forces

- Electrostatics
    - ◊ Polarization
    - ◊ Multipoles, dipoles
    - ◊ VDW Forces
    - ◊ Electrostatic Interactions
- Basic Forces
    - ◊ Electrical non-bonded interactions
    - ◊ bonded, fundamentally QM but treat as springs
    - ◊ Sum up the energy

- Simple Systems First

Overview: Methods for the Generation and Analysis of Macromolecular Simulations

1 Simulation Methods
  - ◊ Potential Functions
  - ◊ Minimization
  - ◊ Molecular Dynamics
  - ◊ Monte Carlo
  - ◊ Simulated Annealing

2 Types of Analysis
  - ◊ liquids: RDFs, Diffusion constants
  - ◊ proteins: RMS, Volumes, Surfaces

- Established Techniques (chemistry, biology, physics)
- Focus on simple systems first (liquids). Then explain how extended to proteins.

Potential Functions

Electric potential, a quick review

- E = electric field = direction that a positive test charge would move
- Force/q = E
- Potential = W/q = work per unit charge = Fx/q = Ex
  - ◊ E = - grad ϕ ; E = ($d\phi/dx$, $d\phi/dy$, $d\phi/dz$)

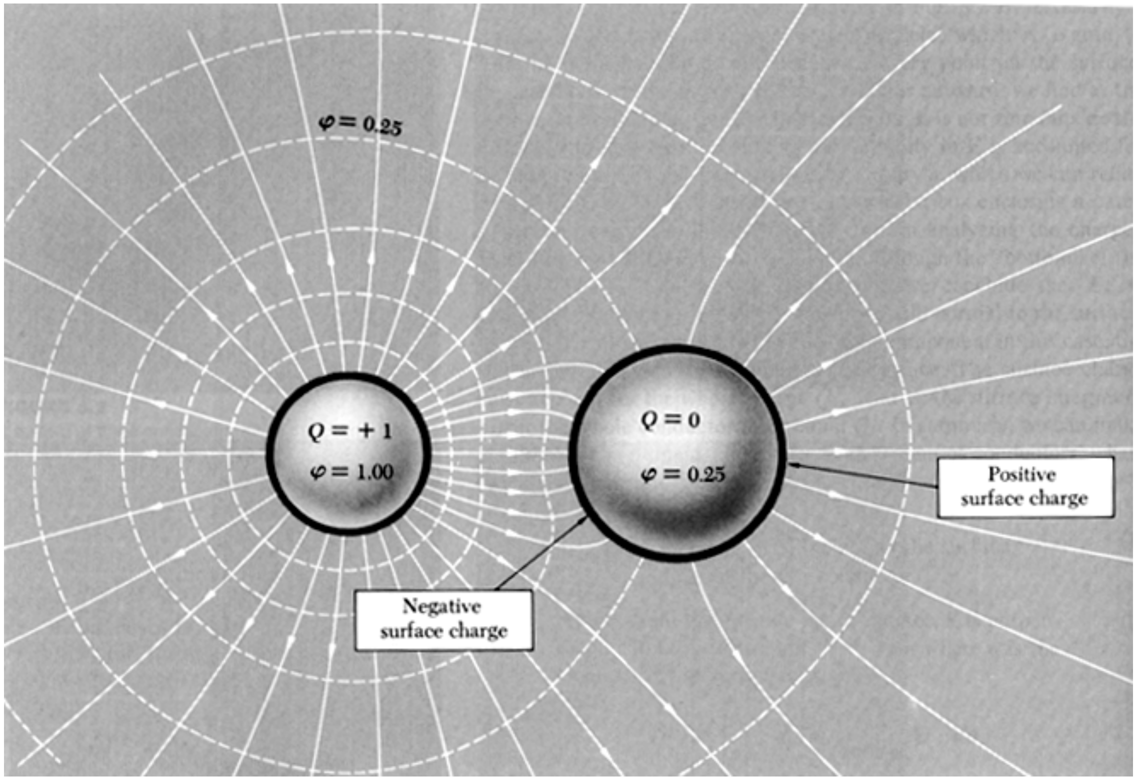
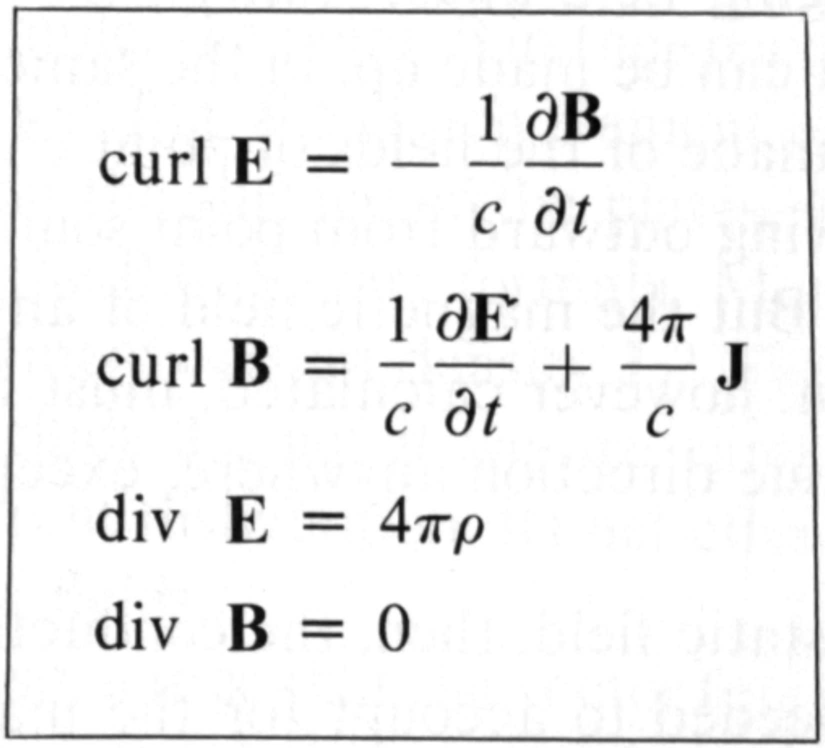


Illustration Credit: Purcell

Maxwell's Equations

- 1st Pair (curl’s)
  - ◊ A changing electric field gives rise to magnetic field that circles around it & vice-versa. Electric Current also gives rise to magnetic field.
    [no discuss here]
- 2nd Pair (div’s)
  - ◊ Relationship of a field to sources
  - ◊ no magnetic monopoles and magnetostatics: div B = 0
    [no discuss here]
- All of Electrostatics in Gauss's Law!!



$$\mathrm{curl}\,\mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}$$

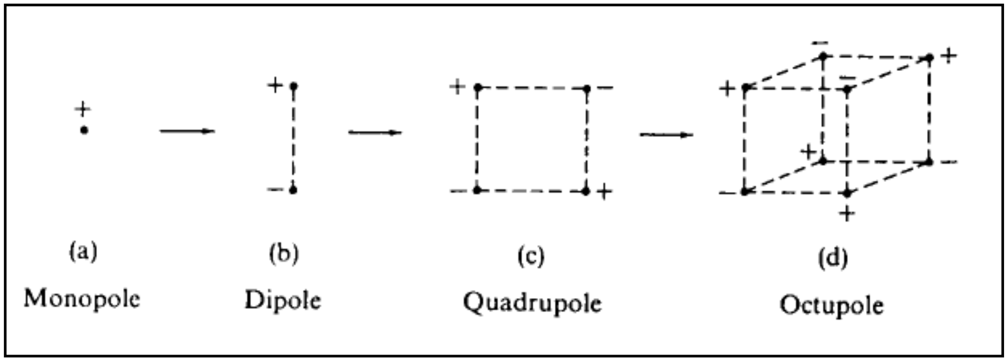
$$\mathrm{curl}\,\mathbf{B} = \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi}{c}\mathbf{J}$$

$$\mathrm{div}\ \mathbf{E} = 4\pi\rho$$

$$\mathrm{div}\ \mathbf{B} = 0$$

cgs (not mks) units above

Multipole Expansion



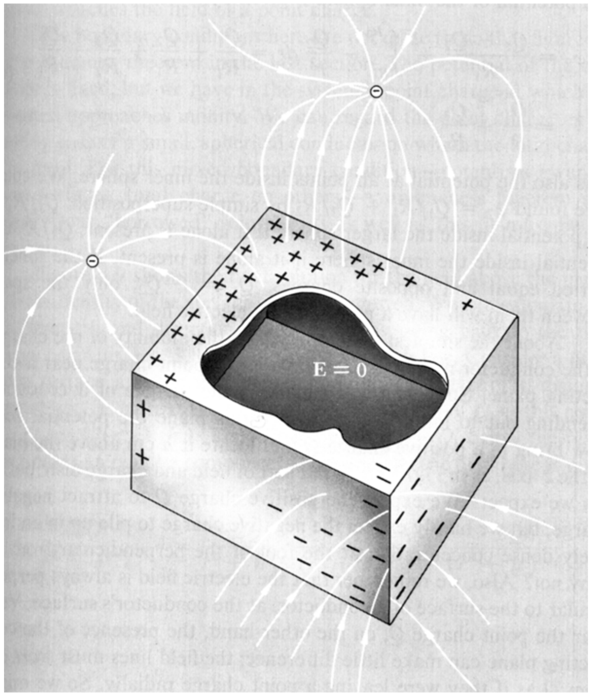
- Routinely done when an atom's charge distribution is replaced by a point charge or a point charge and a dipole
  - ◊ Ignore above dipole here
  - ◊ Harmonic expansion of pot.
- Only applicable far from the charge distribution
  - ◊ Helix Dipole not meaningful close-by
- Terms drop off faster with distance

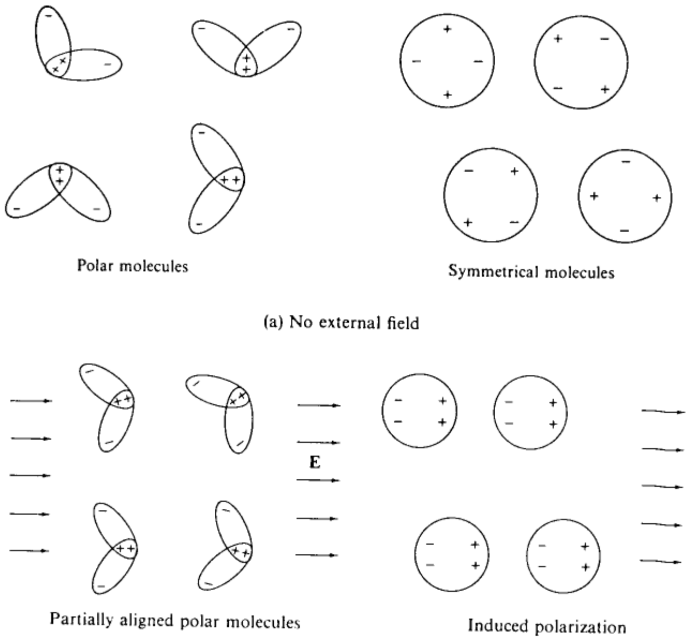
$$\Phi(\mathbf{x}) = \frac{q}{r} + \frac{\mathbf{p} \bullet \mathbf{x}}{r^3} + \frac{1}{2}\sum_{i,j} Q_{ij}\frac{x_i x_j}{r^5} + \cdots$$

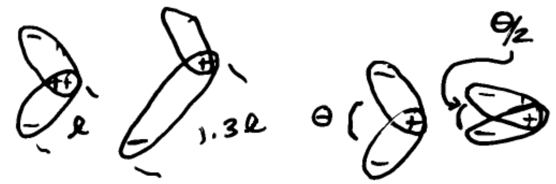
$$\Phi(\mathbf{x}) = \frac{K_1 q}{r} + \frac{K_2 q}{r^2} + \frac{K_3 q}{r^3} + \cdots$$

Replace continuous charge distribution with point moments: charge (monopole) + dipole + quadrupole + octupole + ...

Polarization







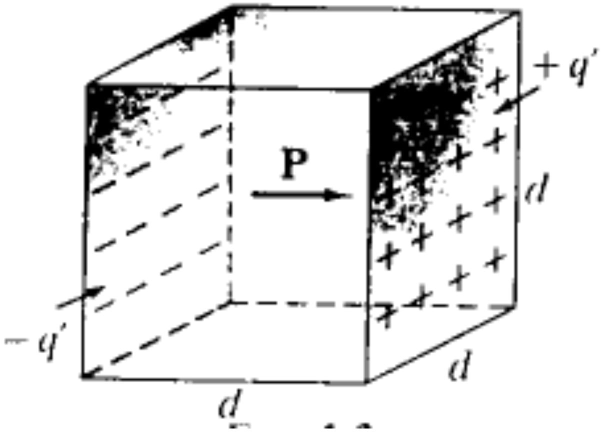
- Charge shifts to resist field
  - ◊ Accomplished perfectly in conductor
    -- surface charge, no field inside
  - ◊ Insulators partially accommodate via induced dipoles
- Induced dipole
  - ◊ charge/ion movement (slowest)
  - ◊ dipole reorient
  - ◊ molecular distort (bond length and angle)
  - ◊ electronic (fastest)

Illustration Credit: Purcell, Marion & Heald

Dielectric const.

- Macro manifestation of polarization
- Values
  (measured in debye)
  - ◊ Air, 1
  - ◊ Water, 80
  - ◊ Paraffin Wax, 2
  - ◊ Methanol, 33
  - ◊ Non-polar protein, 2
  - ◊ Polar protein, 4
- High-frequency
  - ◊ water re-orient, 1ps
  - ◊ bond, angle stretch
  - ◊ electronic, related to index of refraction

$$q'd = Pd^3$$



- $P = \alpha E$
  P = dipole moment per unit volume
- α = electric susceptability
- $\alpha = (\varepsilon - 1)/4\pi$
- ε = dielectric const.
- Effective Field Inside Reduced by Polarization

Molecular Mechanics: Simple electrostatics

atom		ε (kJ/mole)	σ (Å)	charge (electrons)
carbonyl carbon		0.5023	3.7418	0.550
α-carbon	(incorporating 1 hydrogen)	0.2034	4.2140	0.100
β-carbon	(incorporating 3 hydrogens)	0.7581	3.8576	0.000
amide nitrogen		0.9979	2.8509	-0.350
amide hydrogen		0.2085	1.4254	0.250
carbonyl oxygen		0.6660	2.8509	-0.550
water oxygen in interactions with the helix		0.6660	2.8509	-0.834
water hydrogen in interactions with the helix		0.2085	1.4254	0.417
water O in interactions with other waters		0.6367	3.1506	-0.834
water H in interactions with other waters		0.0000	0.0000	0.417

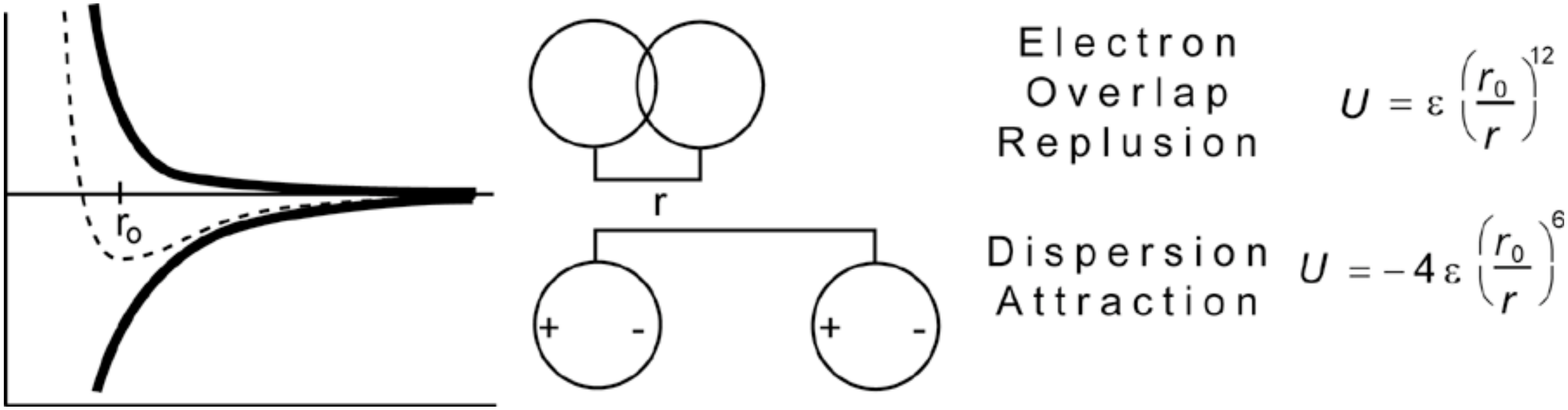
- U = kqQ/r
- Molecular mechanics uses partial unpaired charges with monopole
  - usually no dipole
  - e.g. water has apx. -.8 on O and +.4 on Hs
  - However, normally only use monopoles for unpaired charges (on charged atoms, asp O)
- Longest-range force
  - Truncation? Smoothing

VDW Forces:
Induced dipole-induced dipole

- Too complex to derive induced-dipole-induced dipole formula, but it has essential ingredients of dipole-dipole and dipole-induced dipole calculation, giving an **attractive** $1/r^6$ dependence.
  - ◊ London Forces
- Thus, total dipole cohesive force for molecular system is the sum of three $1/r^6$ terms.
- Repulsive forces result from electron overlap.
  - ◊ Usually modeled as A/r^{12} term. Also one can use exp(-Cr).
- VDW forces: $V(r) = A/r^{12} - B/r^6 = 4\varepsilon((R/r)^{12} - (R/r)^6)$
  - ◊ ε ~ .2 kcal/mole, R ~ 3.5 A, V ~ .1 kcal/mole [favorable]

Packing ~ VDW force

- Longer-range isotropic attractive tail provides general cohesion
- Shorter-ranged repulsion determines detailed geometry of interaction
- Billiard Ball model, WCA Theory



Electron Overlap Replusion

$$U = \varepsilon \left(\frac{r_0}{r}\right)^{12}$$

Dispersion Attraction

$$U = -4\varepsilon \left(\frac{r_0}{r}\right)^{6}$$

H-bonds subsumed by electrostatic interactions

- Naturally arise from partial charges
  - ◊ normally arise from partial charge
- Linear geometry
- Were explicit springs in older models

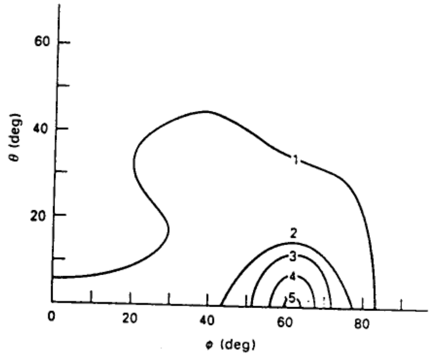


FIGURE 4.4
The geometries of C=O ··· H—N hydrogen bonds observed in crystal structures of small molecules. The definitions of the angles ϕ and θ are illustrated at the *top*, and the relative frequencies of their observed values in intermolecular hydrogen bonds (R. Taylor et al., *J. Amer. Chem. Soc.* 105:5761–5766, 1983) are given by the contours. The angle ϕ measures departures from linearity of the C=O bond and the H atom; the most frequently observed values are in the region of 50°–60°. The angle θ measures the extent to which the H atom lies out of the plane defined by the R, C, and O atoms; the most commonly observed values are in the region of 0°–7°. The lone-pair electrons of the oxygen atom are believed to project at angles of $\phi = 60°$, $\theta = 0°$. The spherical polar coordinate system used here gives a bias toward small values of θ that could be corrected by plotting sin θ.

Illustration Credit: Taylor & Kennard (1984)

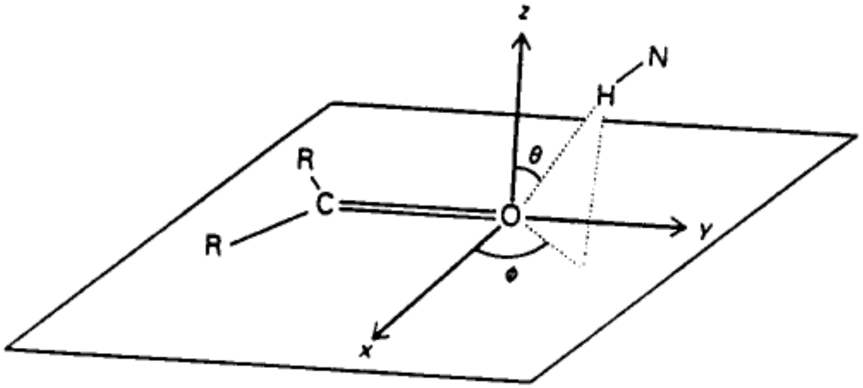


Table 4.7 Lengths of H—N ··· O=C *hydrogen bonds*[a]

	Mean H ··· O Distance for Different Acceptors (Å)		
Donor	Carboxyl[b]	Carboxylate[c]	Amide
$>N-H$[d]	2.002 ± 0.012	1.928 ± 0.012	1.934 ± 0.005
$>N^{+}-H$[e]	1.983 ± 0.055	1.869 ± 0.028	1.858 ± 0.043
NH_4^+	1.916 ± 0.041	1.886 ± 0.018	1.988 ± 0.075
$R-NH_3^+$	1.936 ± 0.014	1.841 ± 0.008	1.891 ± 0.034
$R_2-NH_2^+$	1.887 ± 0.047	1.796 ± 0.014	1.793 ± 0.070
R_3-NH^+		1.722 ± 0.025	1.845 ± 0.014

[a] The N—H distance is generally 1.03 Å; adding this value to the tabulated distances gives the distance between the N and O atoms.

[b] C=O oxygen atom of unionized carboxylic acids and esters.

[c] Oxygen atom of carboxyl anions (—CO_2^-).

[d] Uncharged donor.

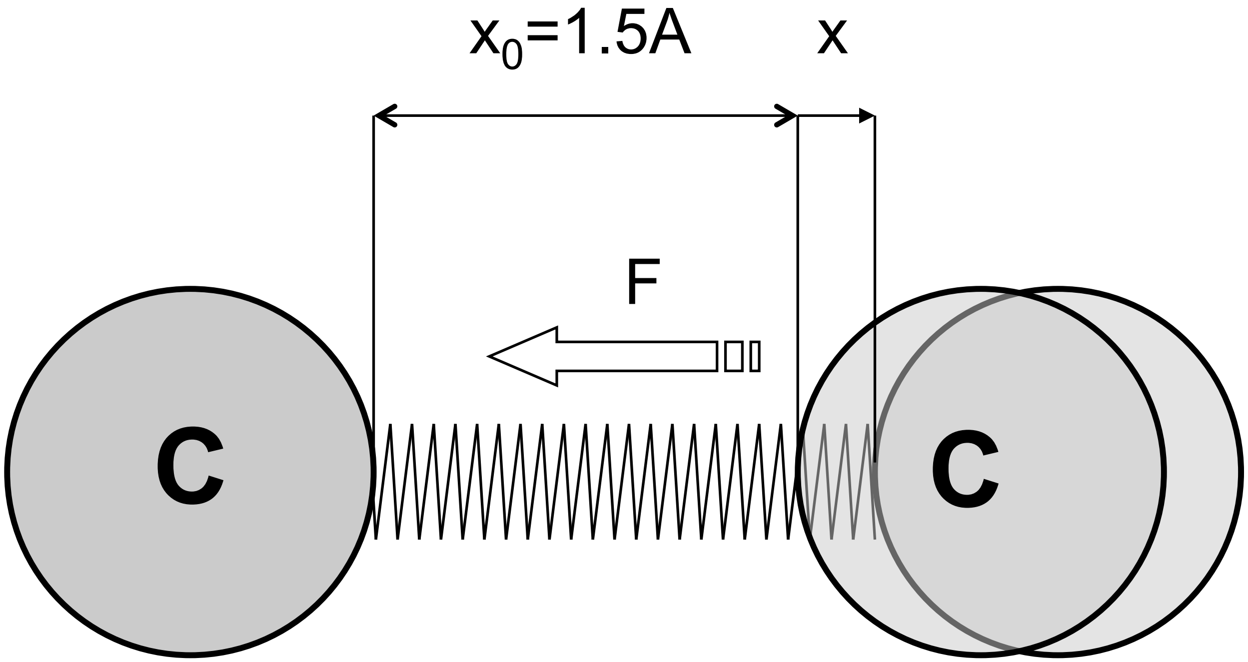
[e] Charged donor with trigonal geometry.

From R. Taylor and O. Kennard, *Acc. Chem. Res.* 17:320–326 (1984).

Hydrophobic interactions

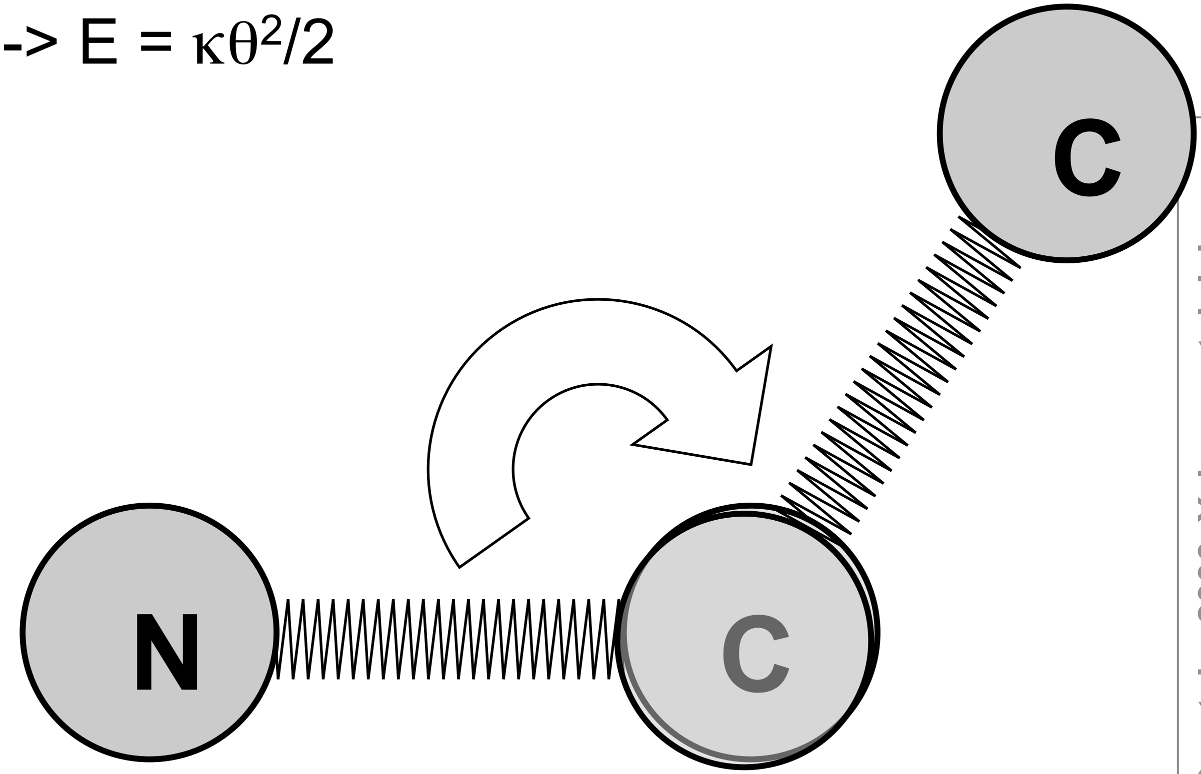
Bond Length Springs

- $F = -kx \rightarrow E = kx^2/2$
- Freq from IR spectroscopy
  - -> w= sqrt(k/m), m = mass => spring const. k
  - k ~ 500 kcal/mole*A^2 (stiff!),
    w corresponds to a period of 10 fs
- Bond length have 2-centers



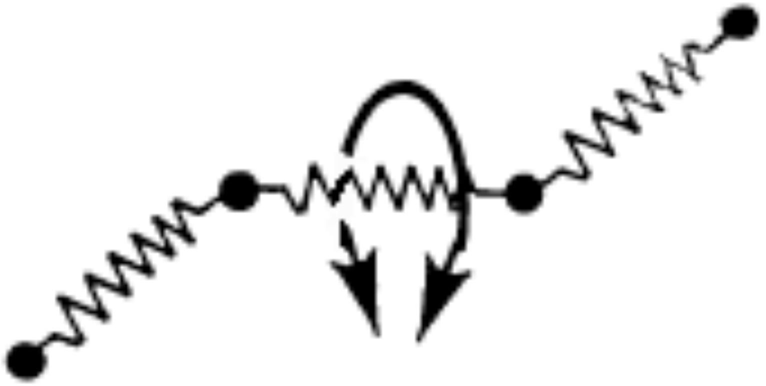
Bond angle, More Springs

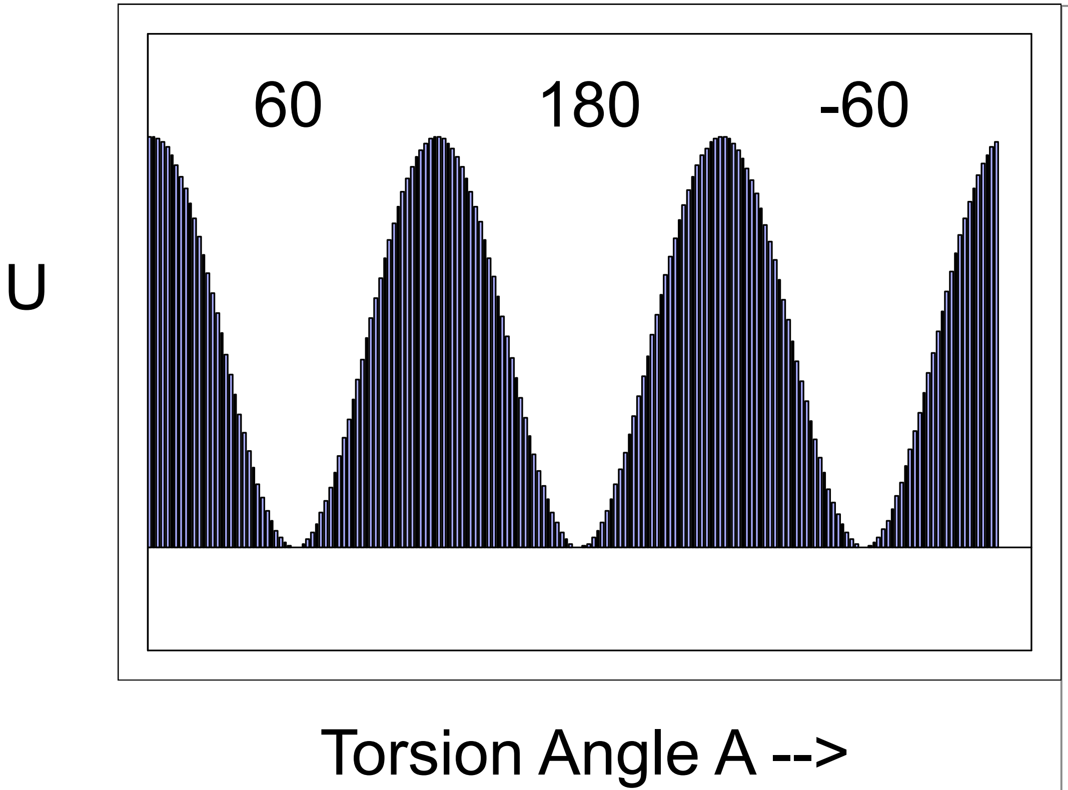
- torque = $\tau = \kappa\theta$ -> E = $\kappa\theta^2/2$
- 3-centers



Torsion angle

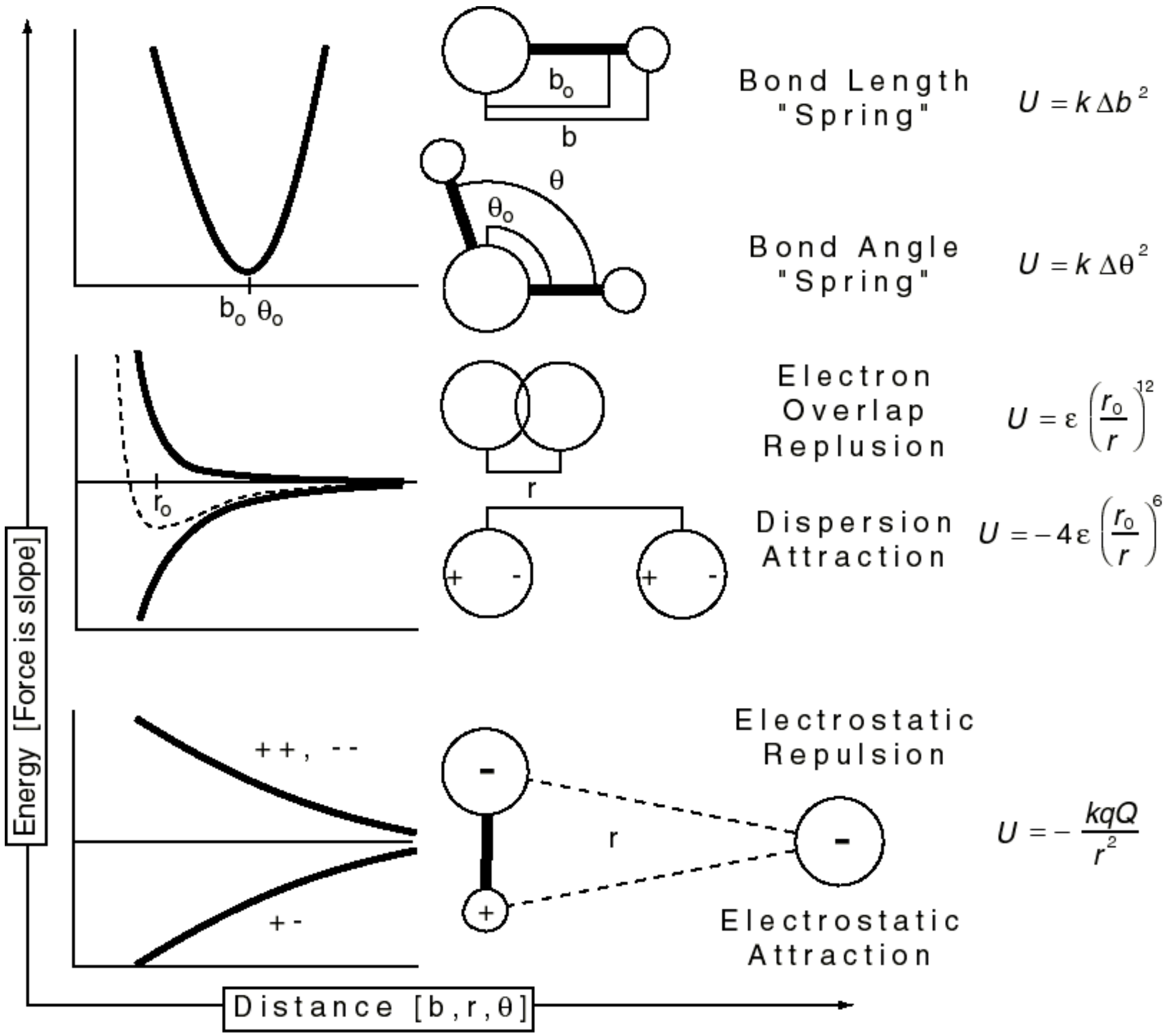
- 4-centers
- $U(A)=K(1-\cos(nA+d))$
  - ◊ $\cos x = 1 + x^2/2 + \ldots$, so minima are quite spring like, but one can hoop between barriers
- K ~ 2 kcal/mole



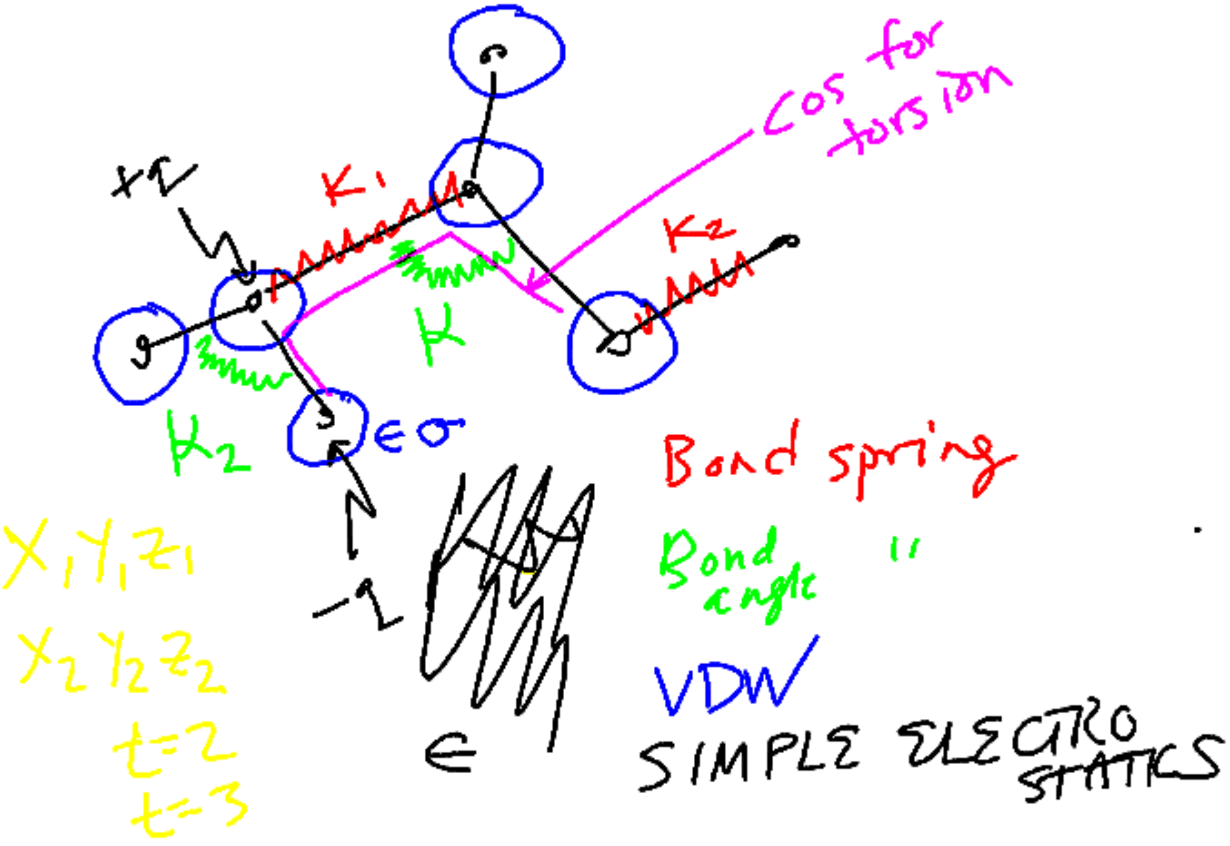


Potential Functions

- Putting it all together
- Springs + Electrical Forces



Summary of the Contributions to the Potential Energy



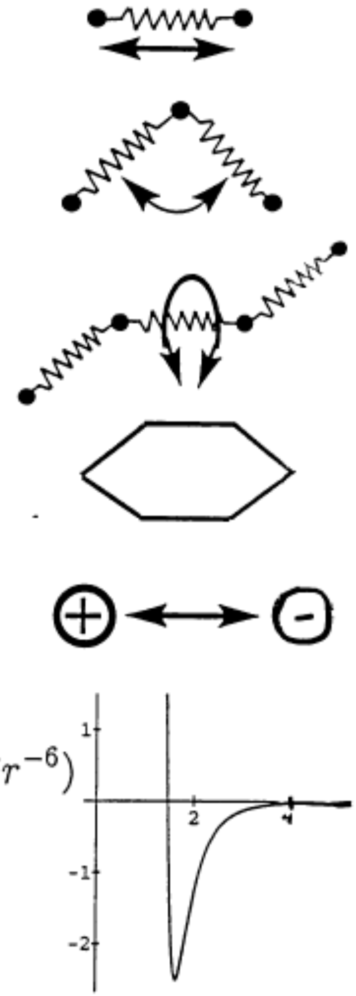
Some of the Simplifications in the Conventional Macromolecular Potential Functions

- Dielectric and polarization effects
- "Motionless" point charges and dipoles
- Bonds as springs

Sum up to get total energy

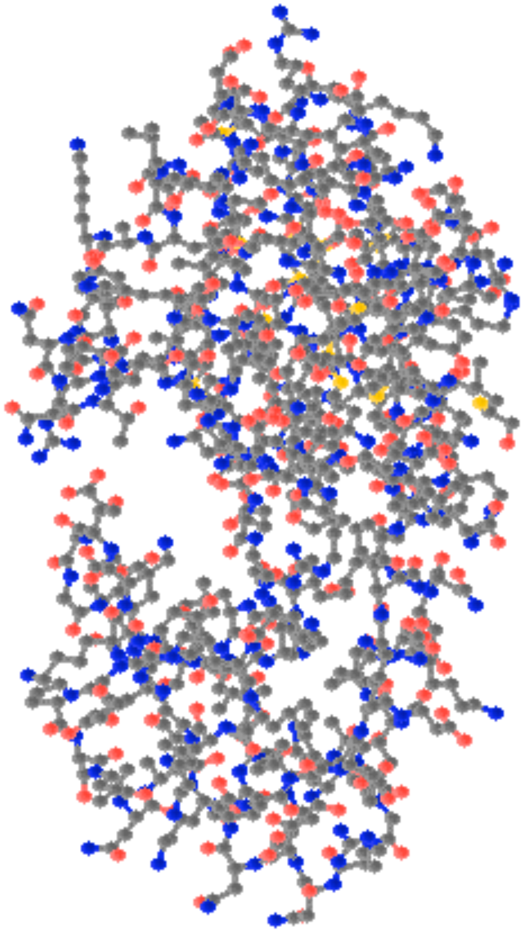
- Each atom is a point mass (m and **x**)
- Sometimes special pseudo-forces: torsions and improper torsions, H-bonds, symmetry.

$$\begin{aligned}
E_{empirical} \;=\;& \sum_{bonds} k_o (b - b_o)^2 \\
&+ \sum_{angles} k_\Phi (\Phi - \Phi_o)^2 \\
&+ \sum_{dihedrals} k_\Psi \cos(n\Psi + \delta) \\
&+ \sum_{chiral,planar\ centers} k_\omega (\omega - \omega_o)^2 \\
&+ \sum_{non-bonded} (Qr^{-1} + Ar^{-12} - Br^{-6}) \\
&+ \sum_{symmetry\ non-bonded} (Qr^{-1} + Ar^{-12} - Br^{-6})
\end{aligned}$$



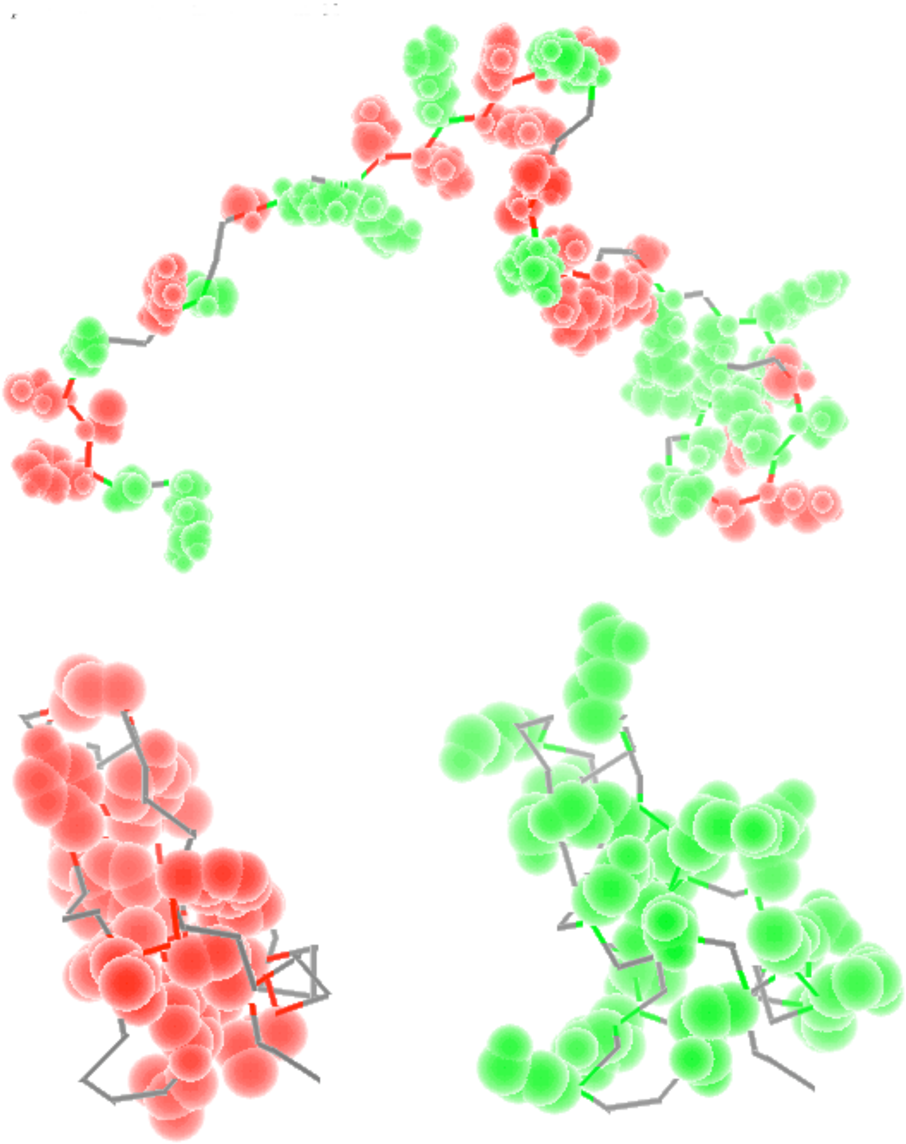
Elaboration on the Basic Protein Model

- Geometry
  - ◊ Start with X, Y, Z’s (coordinates)
  - ◊ Derive Distance, Surface Area, Volume, Axes, Angle, &c
- Energetics
  - ◊ Add Q’s and k’s (Charges for electrical forces, Force Constants for springs)
  - ◊ Derive Potential Function U(x)
- Dynamics
  - ◊ Add m’s and t (mass and time)
  - ◊ Derive Dynamics (v=dx/dt, F = m dv/dt)

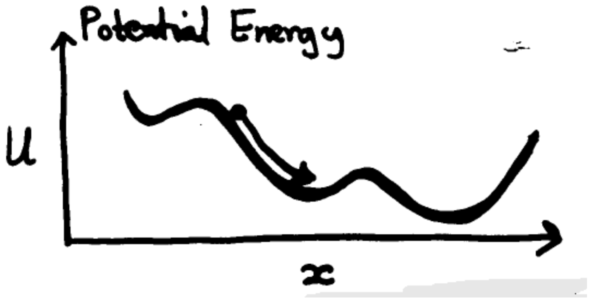


Minimization and Simulation Algorithms for Macromolecules

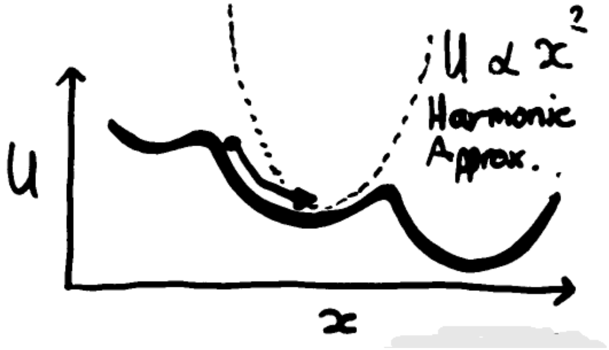
Goal: Model Proteins and Nucleic Acids as Real Physical Molecules



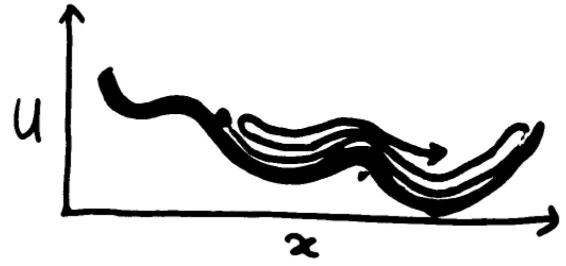
Ways to Move Protein on its Energy Surface



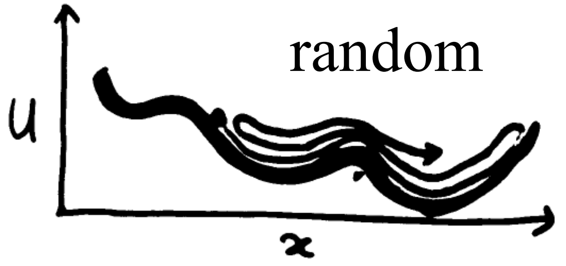
Minimization



Normal Mode Analysis (later?)



Molecular Dynamics (MD)

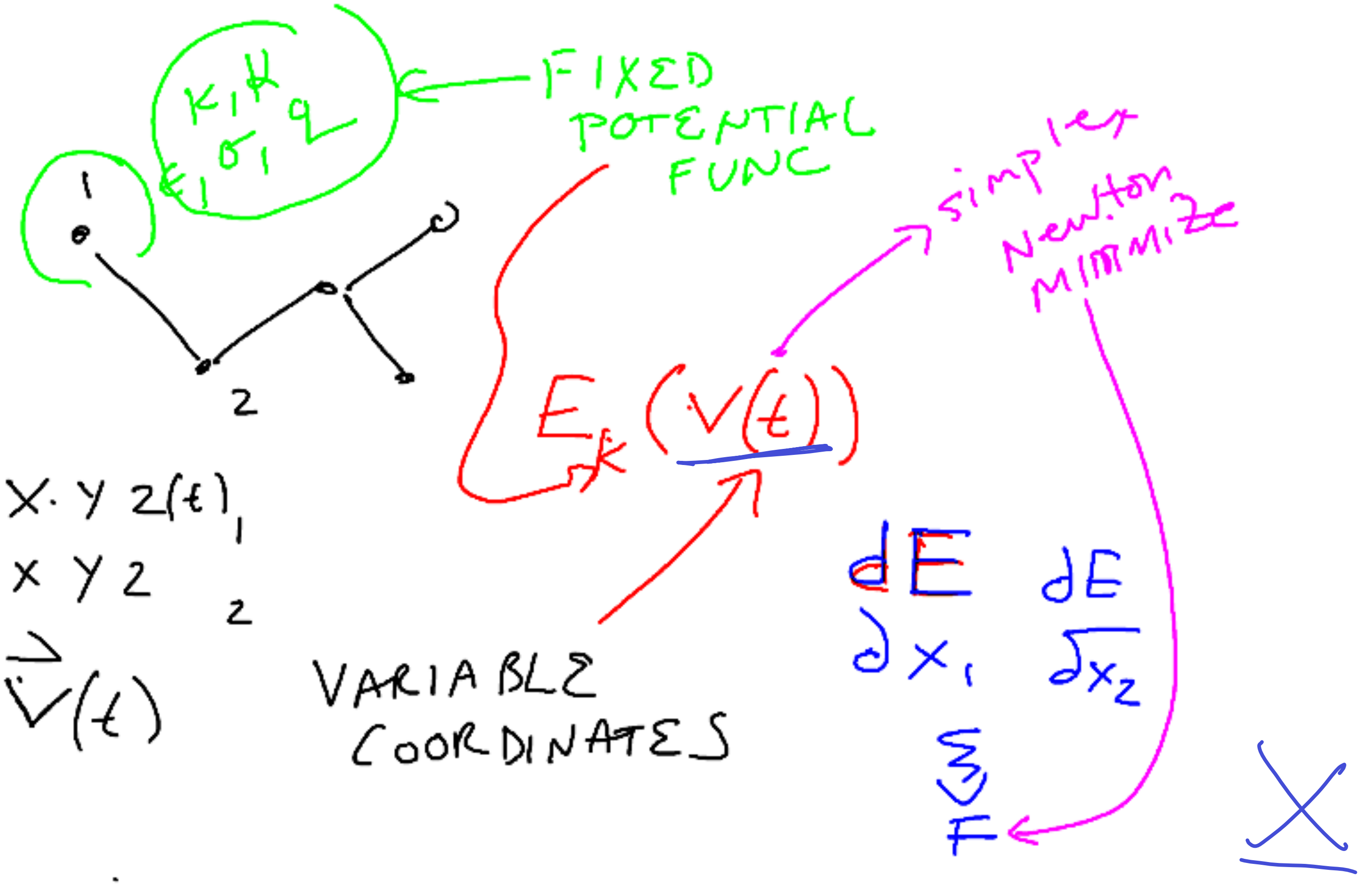


random

Monte Carlo (MC)

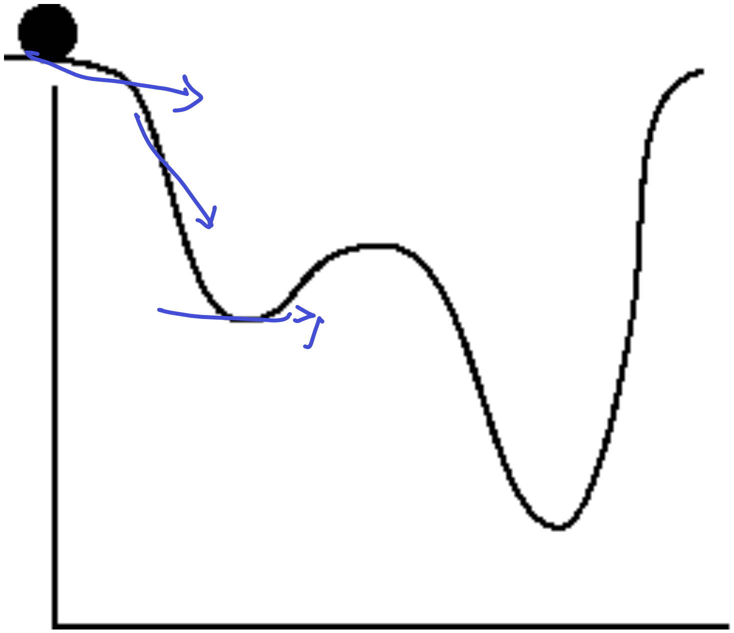
Illustration Credit: M Levitt

Vary the coordinates (XYZs) at a time point t, giving a new Energy E. This can be mimimized with or without derivatives



Steepest Descent Minimization

- Particles on an "energy landscape." Search for minimum energy configuration
  - ◊ Get stuck in local minima
- Steepest descent minimization
  - ◊ Follow gradient of energy straight downhill
  - ◊ i.e. Follow the force:
    step ~ **F** = $-\nabla$ U
    so
    x(t) = **x**(t-1) + a **F**/|**F**|



Multi-dimensional Minimization

- In many dimensions, minimize along lines one at a time
- Ex: U = x^2+$5y^2$, F = (2x,10y)

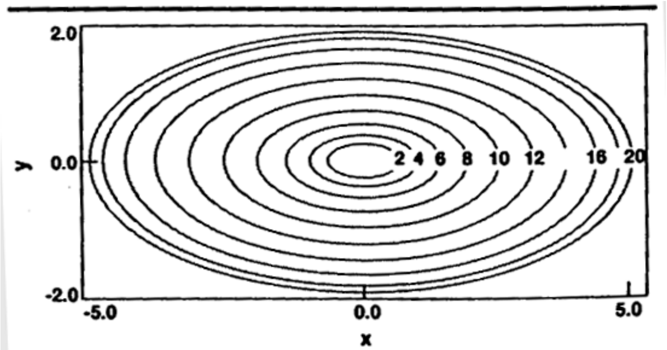


Figure 4–1. Energy Contour Surface of a Simple Function

An energy contour surface for the function $x^2 + 5y^2$. Each contour represents an increase of two arbitrary energy units.

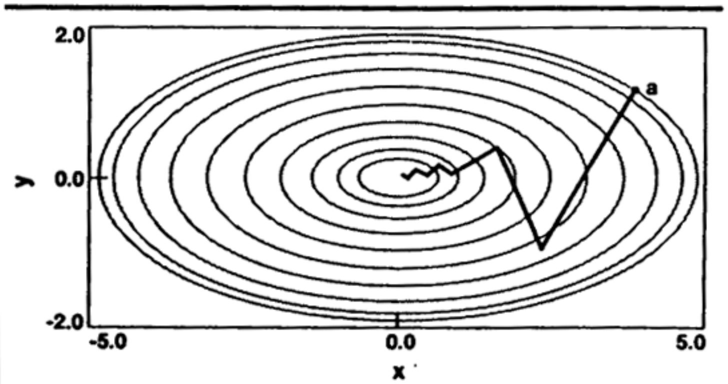


Figure 4–5. Minimization Path following a Steepest-Descents Path without Line Searches

The searching starts from point **a** and converges on the minimum in about 12 iterations. Although the number of iterations is slightly larger than in Figure 4–4, the total minimization is five times faster since, on average, each iteration used only 1.3 function evaluations. Note that, in most applications in molecular mechanics, the function evaluation is the most time-consuming portion of the calculation.

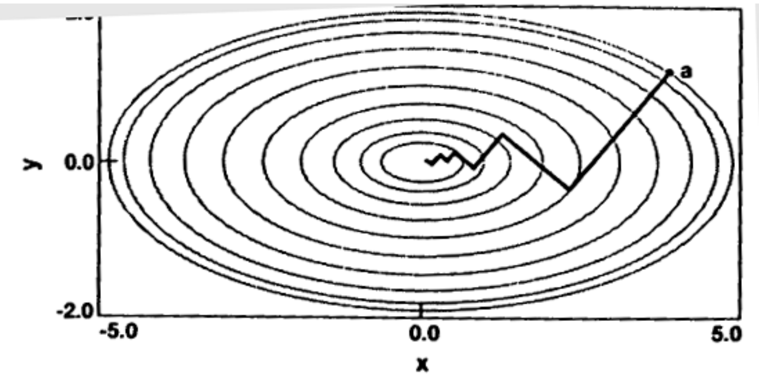


Figure 4–4. Minimization Path following a Steepest-Descents Path

When complete line searches starting from point **a** are used, the minimum is reached in about 12 iterations. Here, where a rigorous line search is carried out, approximately 8 function evaluations are needed for each line search using a quadratic interpolation scheme. Note how steepest descents consistently overshoots the best path to the minimum, resulting in an inefficient, oscillating trajectory.

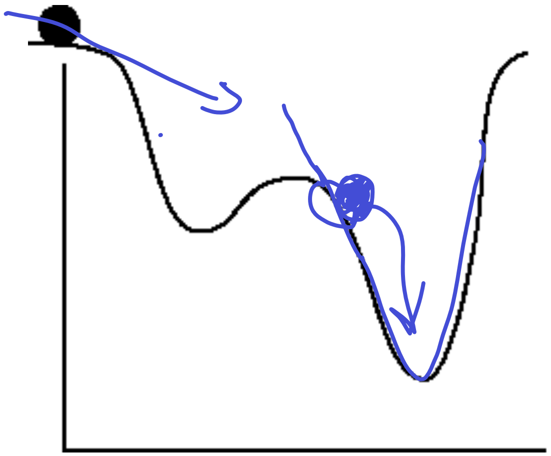
Illustration Credit: Biosym, discover manual

Other Minimization Methods

- Simplex, grid search
  - ◊ no derivatives
- Conjugate gradient

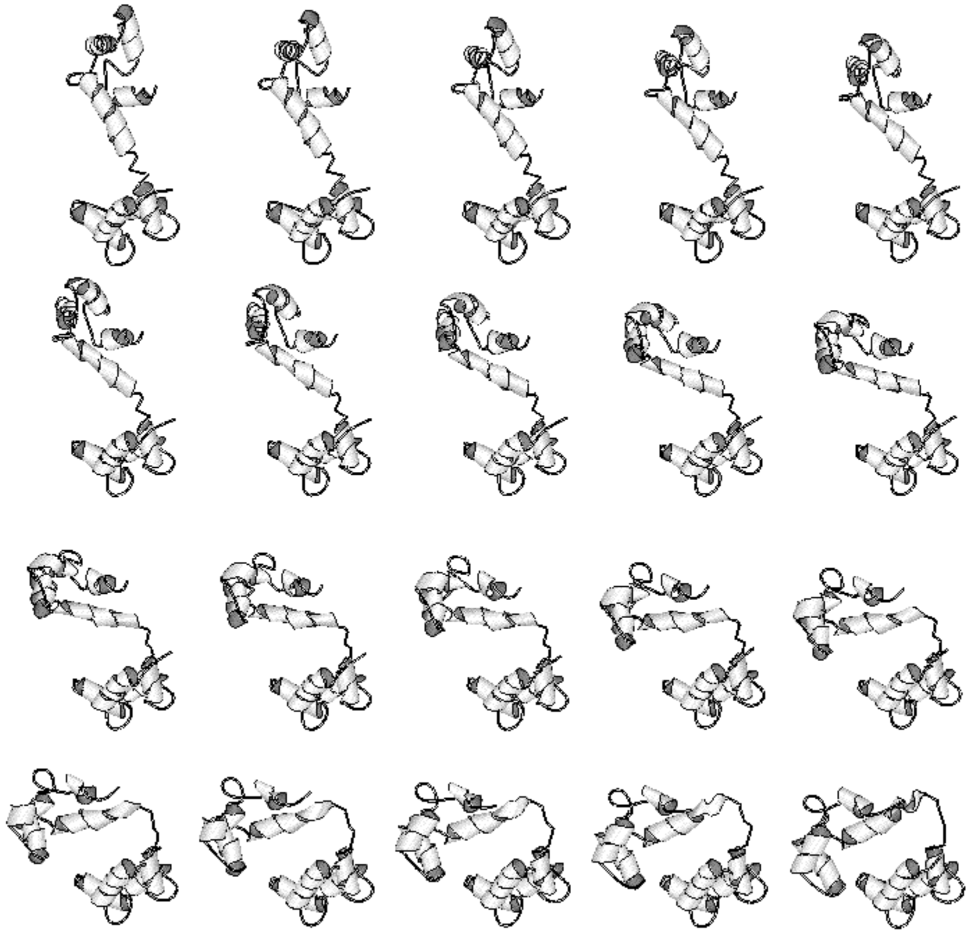
  step ~ F(t) **-** b**F**(t-1)
  - ◊ partial 2nd derivative
- Newton-Raphson
  - ◊ using 2nd derivative, find minimum assuming it is parabolic
  - ◊ $V = ax^2 + bx + c$
  - ◊ $V' = 2ax + b$ & $V'' = 2a$
  - ◊ $V' = 0 \rightarrow x^* = -b/2a$

- Problem is that get stuck in local minima
- Steepest descent, least clever but robust, slow at end
- Newton-Raphson faster but 2nd deriv. can be fooled by harmonic assumption
- Recipe: steepest descent 1st, then Newton-raph. (or conj. grad.)



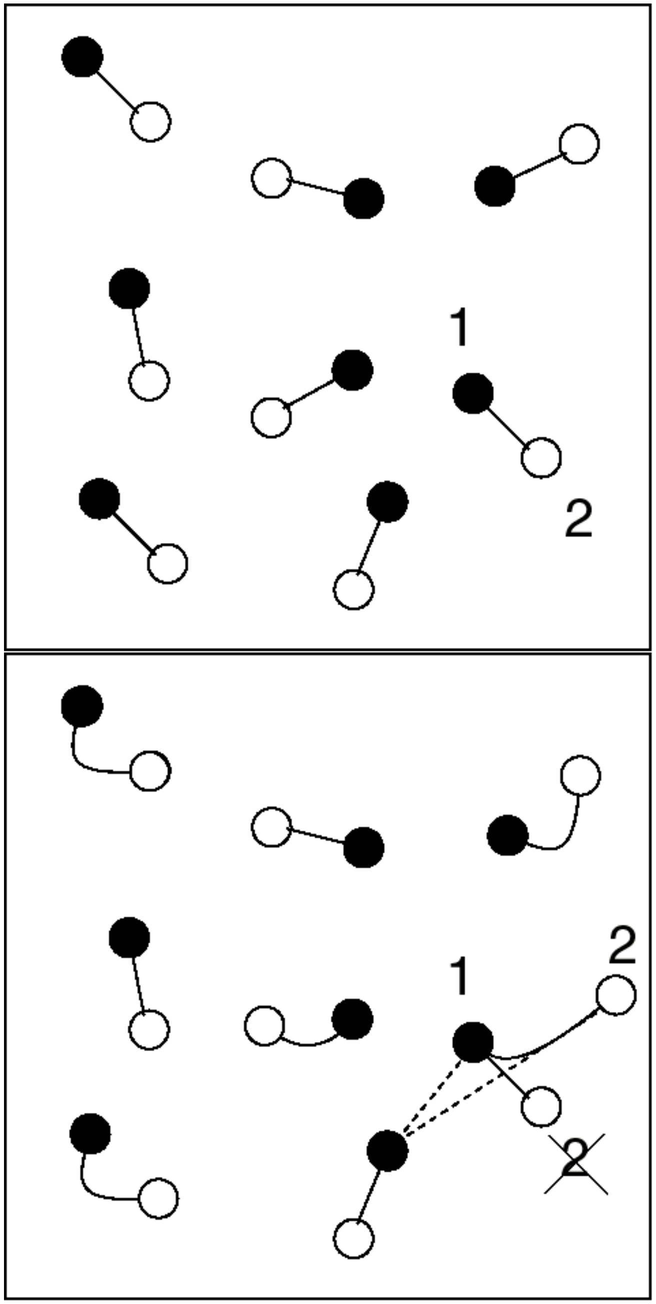
Adiabatic mapping

- Interpolate then minimize
  - ◊ Gives apx. energy (H) landscape through a barrier
  - ◊ can sort of estimate transition rate rate = (kT/h) exp (-dG/kT)
  - ◊ Used for ring flips, hinge motions



Molecular Dynamics

- Give each atoms a velocity.
  - ◊ If no forces, new position of atom (at t + dt) would be determined only by velocity
    x(t+dt) = **x**(t) + **v** dt
- Forces change the velocity, complicating things immensely
  - ◊ **F** = d**p/dt** = m d**v**/dt



Molecular Dynamics (cont)

- On computer make very small steps so force is nearly constant and velocity change can be calculated (uniform a)

[Avg. **v** over Δt] = (**v** + Δ**v**/2)

- Trivial to update positions:

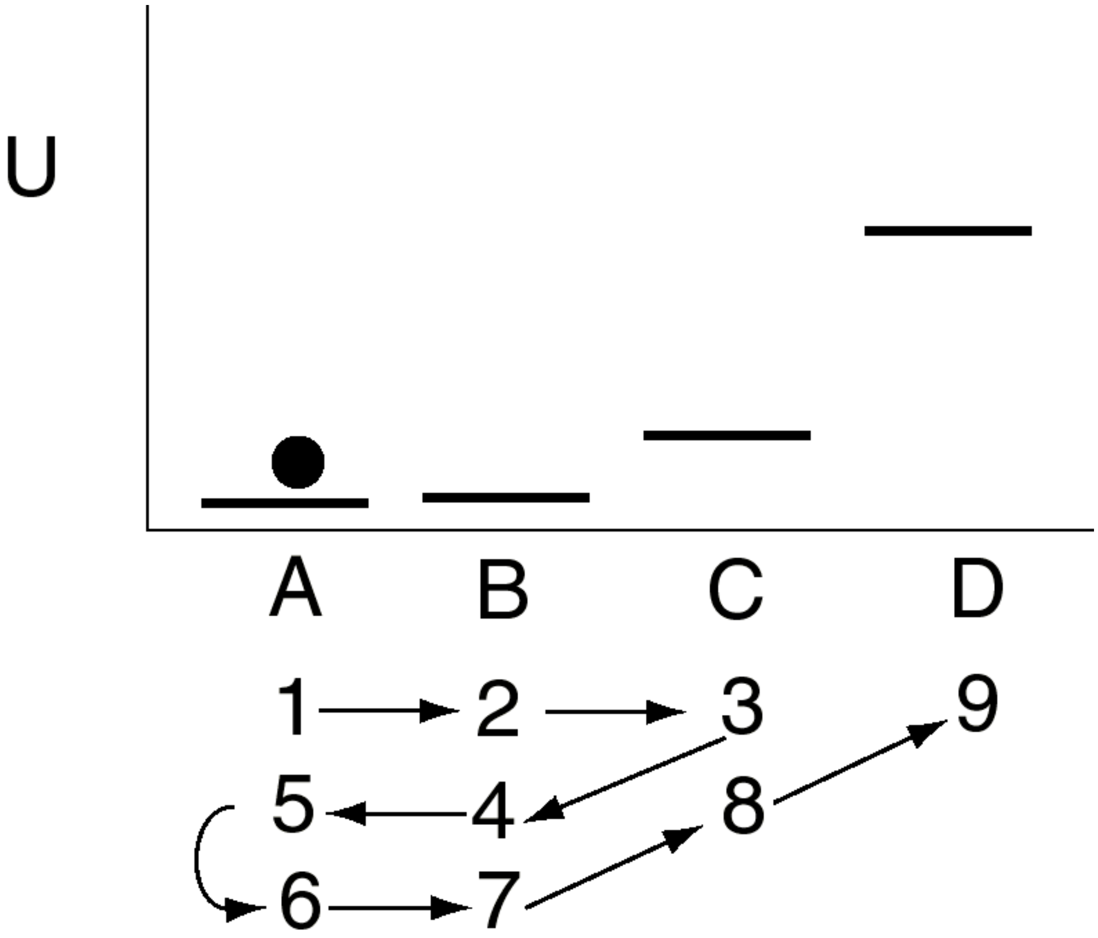
$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \left(\mathbf{v} + \frac{\Delta\mathbf{v}}{2}\right)\Delta t$$
$$= \mathbf{x}(t) + \mathbf{v}\Delta t + \frac{\mathbf{F}}{2m}\Delta t^2$$

- Step must be very small
  - ◊ Δt ~ 1fs
    (atom moves 1/500 of its diameter)
  - ◊ This is why you need fast computers
- Actual integration schemes slightly more complicated
  - ◊ Verlet (explicit half-step)
  - ◊ Beeman, Gear
    (higher order terms than acceleration)

Phase Space Walk

- Trajectories of all the particles traverses space of all possible configuration and velocity states (phase space)
- Ergodic Assumption:
  Eventually, trajectory visits every state in phase space
- Boltzmann weighting:
  Throughout, trajectory samples states fairly in terms of system's energy levels
  - ◊ More time in low-U than high-U states
  - ◊ Probability of being in a state ~ exp(-U/kT)
- Consequently, statistics (average properties) over trajectory are thermodynamically correct

Example Phase Space Walk



$(X) = 3X_A + 3X_B + 2X_C + 1X_D$

$(U) = 6U_{AB} + 2U_C + 1U_D$

Monte Carlo

- Other ways than MD to sample states fairly and compute correctly weighted averages? Yes, using Monte Carlo calculations.
- Basic Idea: Move through states randomly, accepting or rejecting them so one gets a correct "Boltzmann weighting"

MCMC

- Formalism:
  - ◊ System described by a probability distribution $\rho(n)$ for it to be in each state n
  - ◊ Random ("Markov") process π operates on the system and changes distribution amongst states to $\pi\rho(n)$
  - ◊ At equilbrium original distribution and new distribution have to be same as Boltzmann distribution

LOCAL CONS.

DETAILED BALANCE

$$\pi\rho(n) = \rho(n) = \frac{1}{Z}\exp\left(\frac{-U(n)}{kT}\right)$$

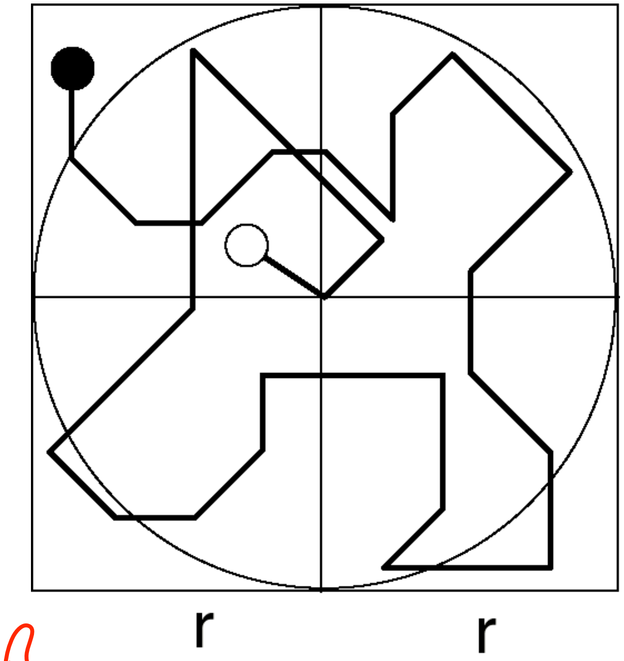
→ GLOBAL PROP

Monte Carlo (cont)

- Metropolis Rule (for specifying π)
  1. Make a random move to a particle and calculate the energy change dU
  2. dU < 0 –> accept the move
  3. Otherwise, compute a random number R between 0 and 1: R < ~ exp(-U/kT) –>
     accept the move
     otherwise –>
     reject the move

LATER

- "Fun" example of MC Integration
  - ◊ Particle in empty box of side 2r (energy of all states same)
  - ◊ π = 6 x [Fraction of times particles is within r of center]

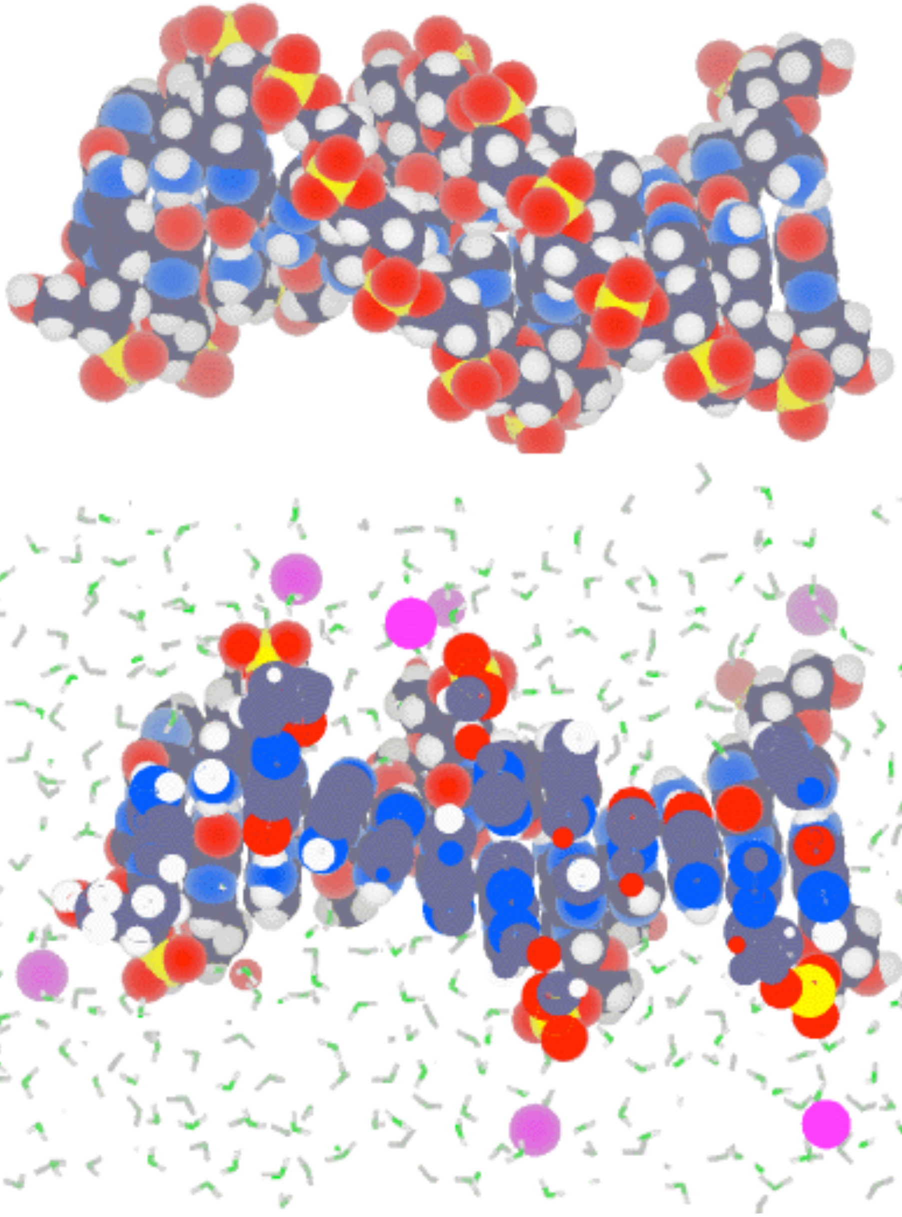


MC vs/+ MD

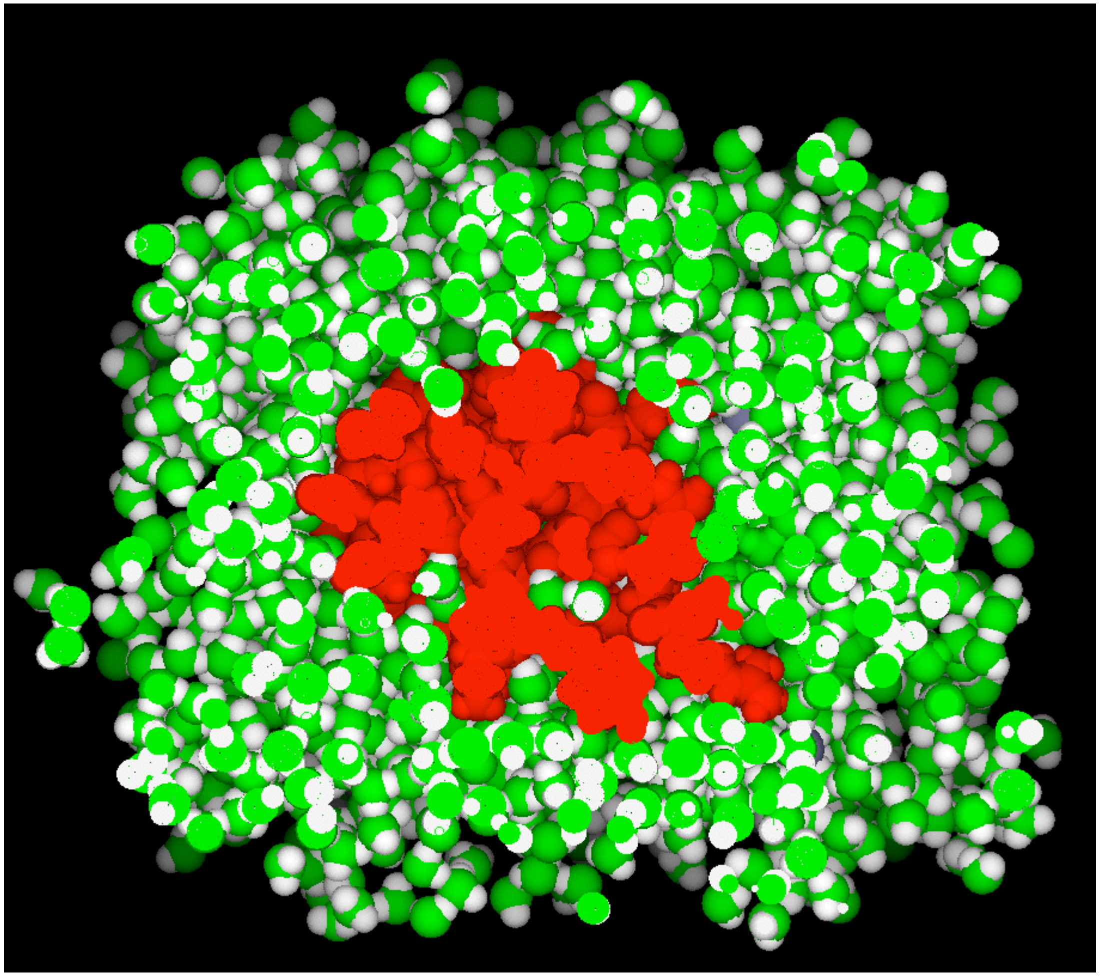
- MD usually used for proteins. Difficult to make moves with complicated chain.
- MC often used for liquids. Can be made into a very efficient sampler.
- Hybrid approaches (Brownian dynamics)
- Simulated Annealing. Heat simulation up to high T then gradually cool and minimize to find global minimum.

Practical Aspects of Simulation

Typical Systems: DNA + Water



Typical Systems: Protein + Water



Practical Aspects: simulation cycle I

- Divide atoms into types (e.g. alpha carbon except for Gly, carbonyl oxygen)
- Initially
  - Associate each atom with a mass and a point charge
  - Give each atom an initial velocity
- Calculate Potential
- Calculating non-bonded interactions take up all the time

CUTOFF

  - Electrostatics hardest since longest ranged
  - Neighbor lists

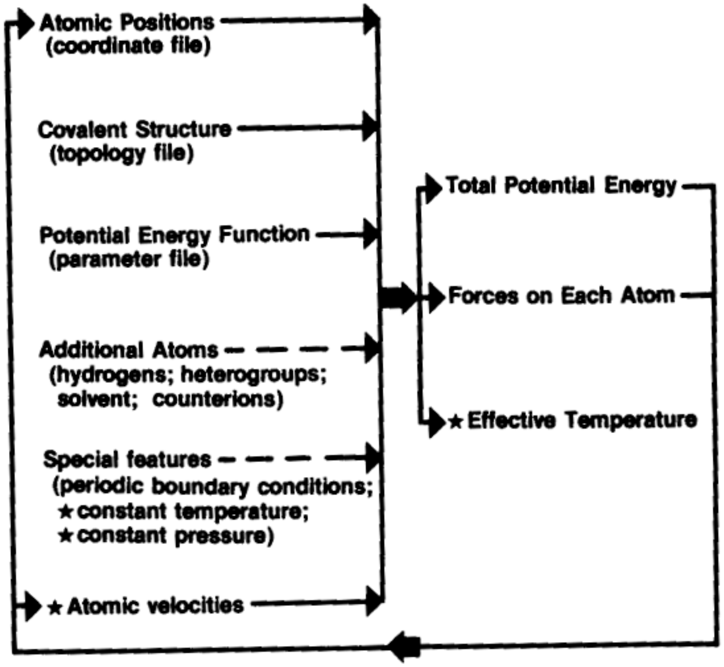


Fig. 4.1. Schematic flow chart of algorithms for energy minimization and molecular dynamics. Features which apply only to molecular dynamics are indicated by asterisks. Dashed lines indicate optional input. Each cycle of energy minimization represents a step in conformation space, while each cycle of molecular dynamics represents a step in time.

Illustration Credit: McCammon & Harvey (1987)

Practical Aspects: simulation cycle II

- Update Positions with MD equations, then recalculate potential and continue
- Momentum conservation
- Energy Conserved in NVE ensemble
- Hydrophobic interaction naturally arises from water behavior

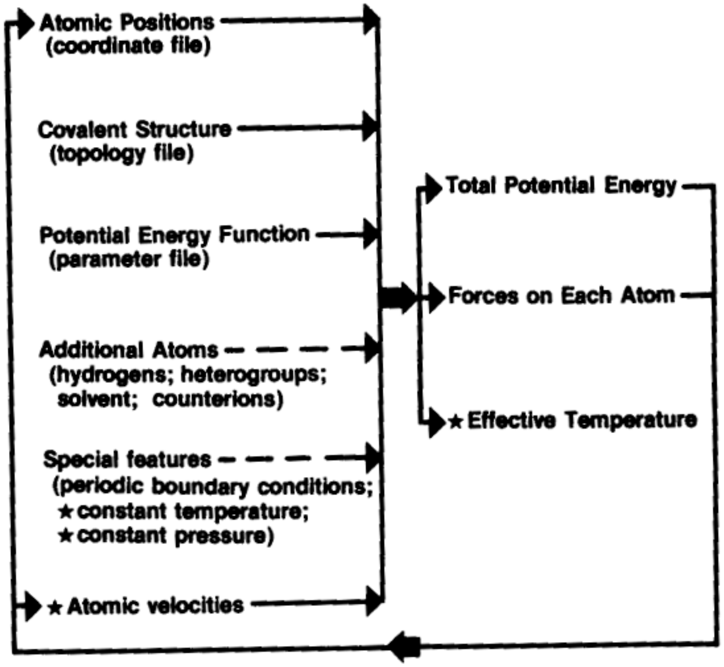
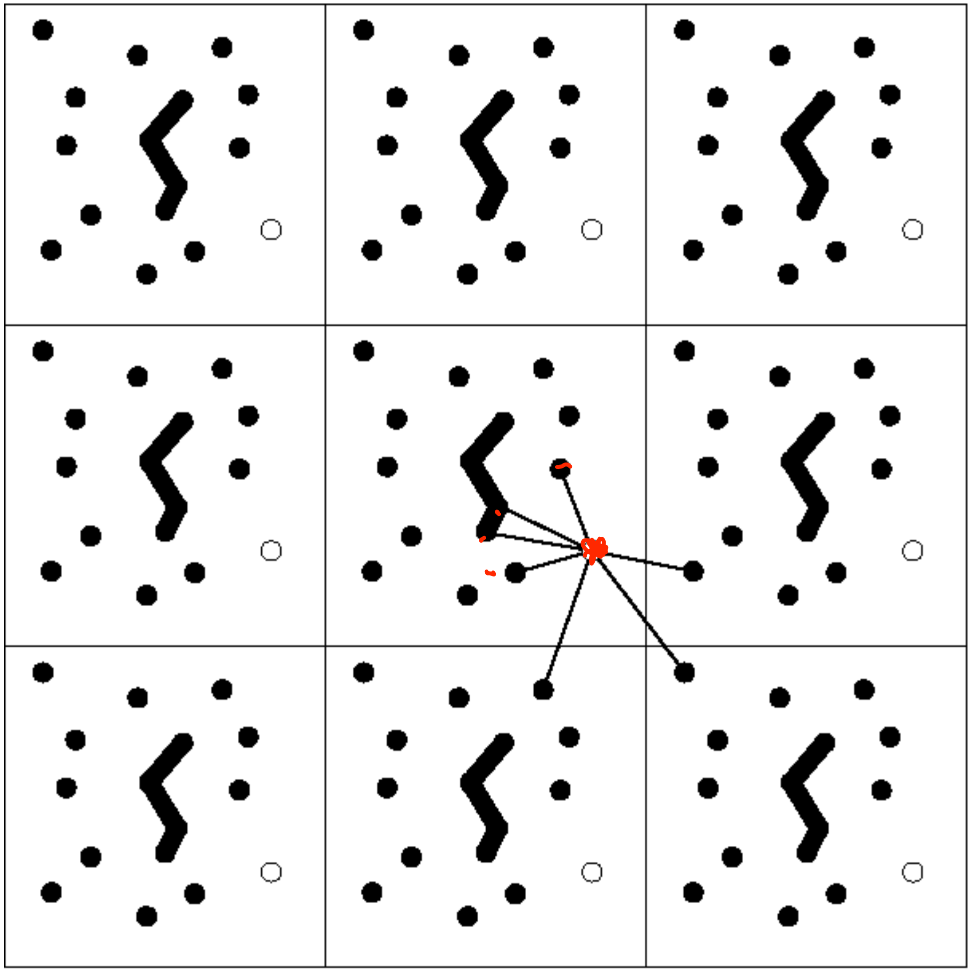


Fig. 4.1. Schematic flow chart of algorithms for energy minimization and molecular dynamics. Features which apply only to molecular dynamics are indicated by asterisks. Dashed lines indicate optional input. Each cycle of energy minimization represents a step in conformation space, while each cycle of molecular dynamics represents a step in time.

Illustration Credit: McCammon & Harvey (1987)

Periodic Boundary Conditions

- Make simulation system seem larger than it is
- Ewald Summation for electrostatics (Fourier transform)



End of class M8
[2006,11.27]
Start of class M9
[2006,11.29]

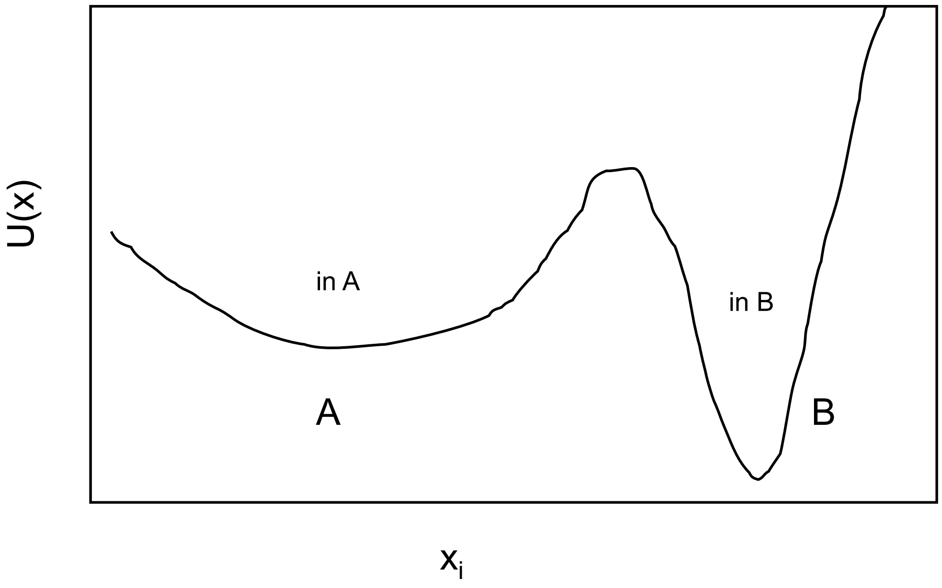
Analysis: What can be Calculated from Simulation?

Average over simulation

- Deceptive Instantaneous Snapshots (almost anything can happen)
- Simple thermodynamic averages
  - ◊ Average potential energy <U>
  - ◊ T ~ < Kinetic Energy > = ½ m < v^2 >
- Some quantities fixed, some fluctuate in different ensembles
  - ◊ NVE protein MD (“microcanonical”)
  - ◊ NVT liquid MC (“canonical”)
  - ◊ NPT more like the real world

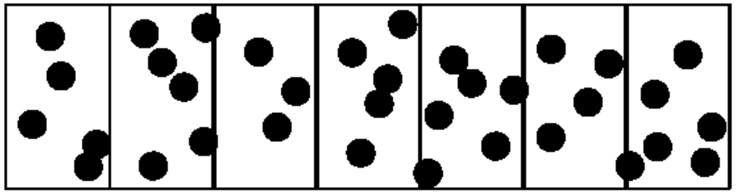
Energy and Entropy

- Energy
  - ◊ At each point i (with coordinates x_i) on the pot. energy surface there is a well-defined “energy” $U(x_i)$
- Probability of occurrence
  - ◊ $P_i = \exp(-U_i/kT)/Q$
  - ◊ The boltzmann distribution
  - ◊ Q = Sum over all P_i , to normalize probabilities to 1



- Entropy
  - ◊ $S(A) = k \sum (P_i \ln P_i)$, where the sum is over points i in A
- Free Energy
  - ◊ $G(A) = U(A) - TS(A)$
- Entropy and Free Energy are only defined for distinctly diff. “states” -- e.g. A (“unfolded”)and B (“folded”)
  - ◊ State B has a lower U and its minimum is more probable than State A
  - ◊ However, state A has a broader minimum that can be occupied in more ways
- Relative Prob
  - ◊ $P(A)/P(B) = \dfrac{\exp(-G(A)/kT)}{\exp (G(B)/kT)}$

Number Density



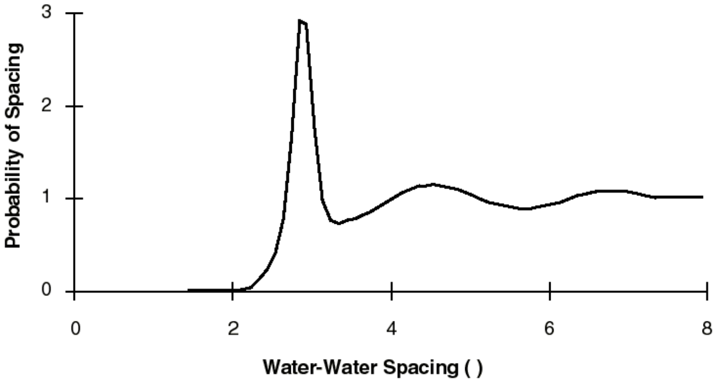
Observed	5	6	3	5	6	4	6
Expected	5	5	5	5	5	5	6

= Number of atoms per unit volume averaged over simulation divided by the number you expect to have in the same volume of an ideal “gas”

Spatially average over all directions gives

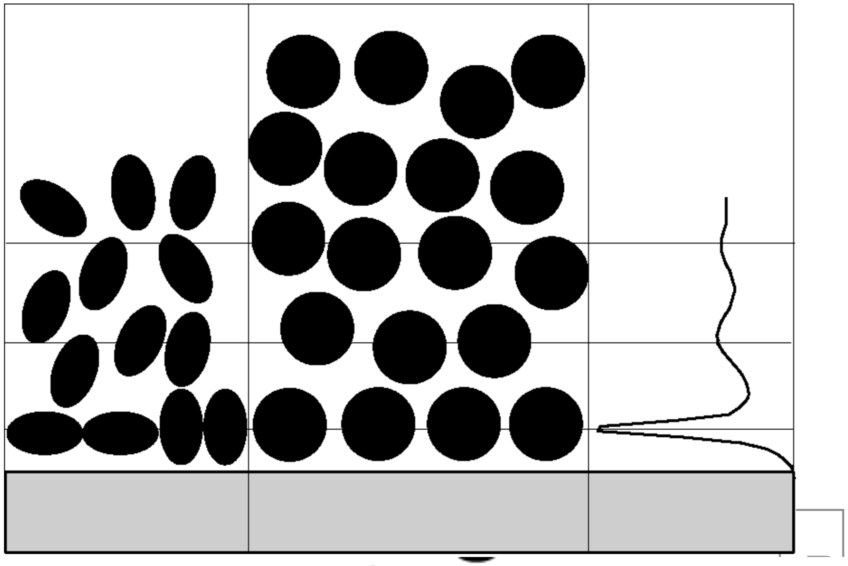
1D RDF =

$$\frac{[\text{ Avg. Num. Neighbors at } r \quad]}{[\text{Expected Num. Neighbors at } r\ ]}$$

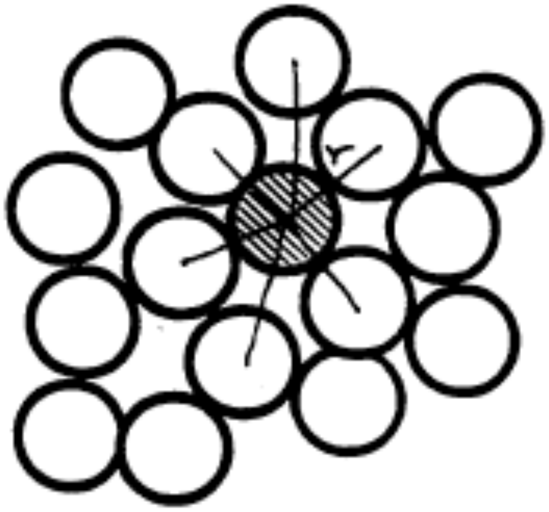


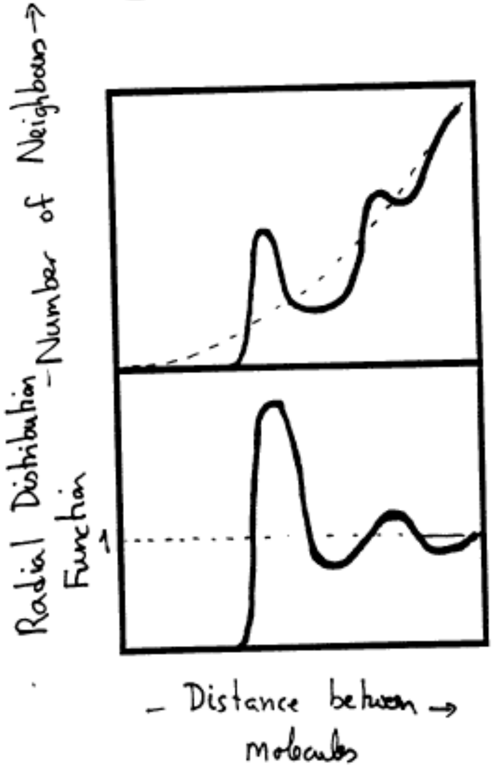
“at r” means contained in a thin shell of thickness dr and radius r.

Number Density (cont)



- Advantages: Intuitive, Relates to scattering expts
- D/A: Not applicable to real proteins
  - ◊ 1D RDF not structural
  - ◊ 2D proj. only useful with "toy" systems
- Number densities measure spatial correlations, not packing
  - ◊ Low value does not imply cavities
  - ◊ Complicated by asymmetric molecules
  - ◊ How things pack and fit is property of instantaneous structure - not average





Measurement of Dynamic Quantities I

- The time-course of a relevant variable is characterized by

(1) Amplitude (or magnitude), usually characterized by an RMS value

$$R = \mathrm{sqrt}[\ < (a(t) - \langle a(t)\rangle)^2 >\ ]$$
$$R = \mathrm{sqrt}[\ < a(t)^2 - 2a(t)\langle a(t)\rangle + \langle a(t)\rangle^2 >\ ]$$
$$R = \mathrm{sqrt}[\ < a(t)^2> - \langle a(t)\rangle^2\ ]$$

  - similar to SD
  - fluctuation

- Relevant variables include bond length, solvent molecule position, H-bond angle, torsion angle

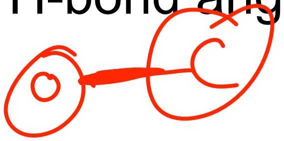
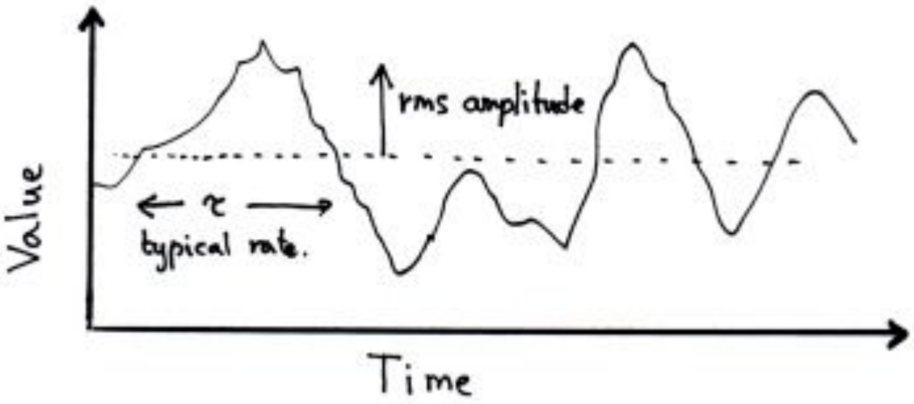


Illustration from M Levitt, Stanford University



Measurement of Dynamic Quantities II

- The time-course of a relevant variable is characterized by

(2) Rate or time-constant

  - ◊ Time Correlation function
  - ◊ $C_A(t) = \langle A(s)A(t+s)\rangle = \langle A(0)A(t)\rangle$ [averaging over all s]
  - ◊ Correlation usually exponentially decays with time t
  - ◊ decay constant is given by the integral of C(t) from t=0 to t=infinity

- Relevant variables include bond length, solvent molecule position, H-bond angle, torsion angle

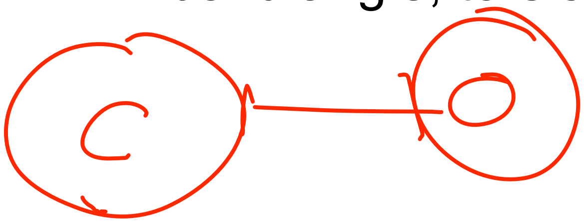
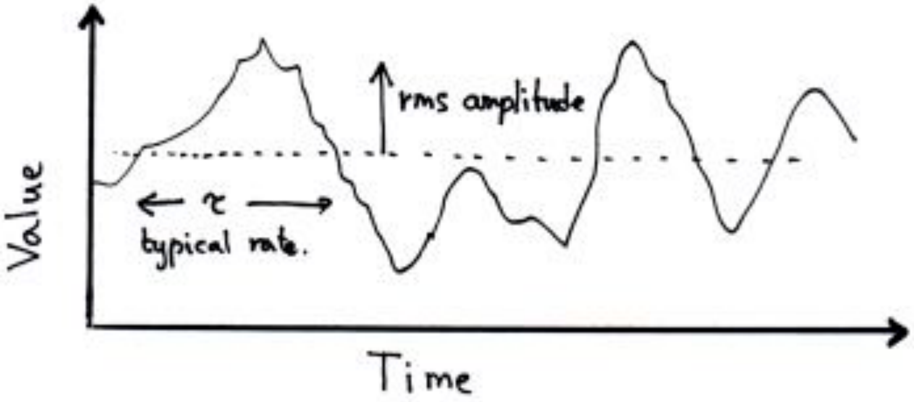


Illustration from M Levitt, Stanford University



D & RMS

- Diffusion constant
  - ◊ Measures average rate of increase in variance of position of the particles
  - ◊ Suitable for liquids, not really for proteins

$$D = \frac{\langle \Delta r^2 \rangle}{6\Delta t}$$

- RMS more suitable to proteins

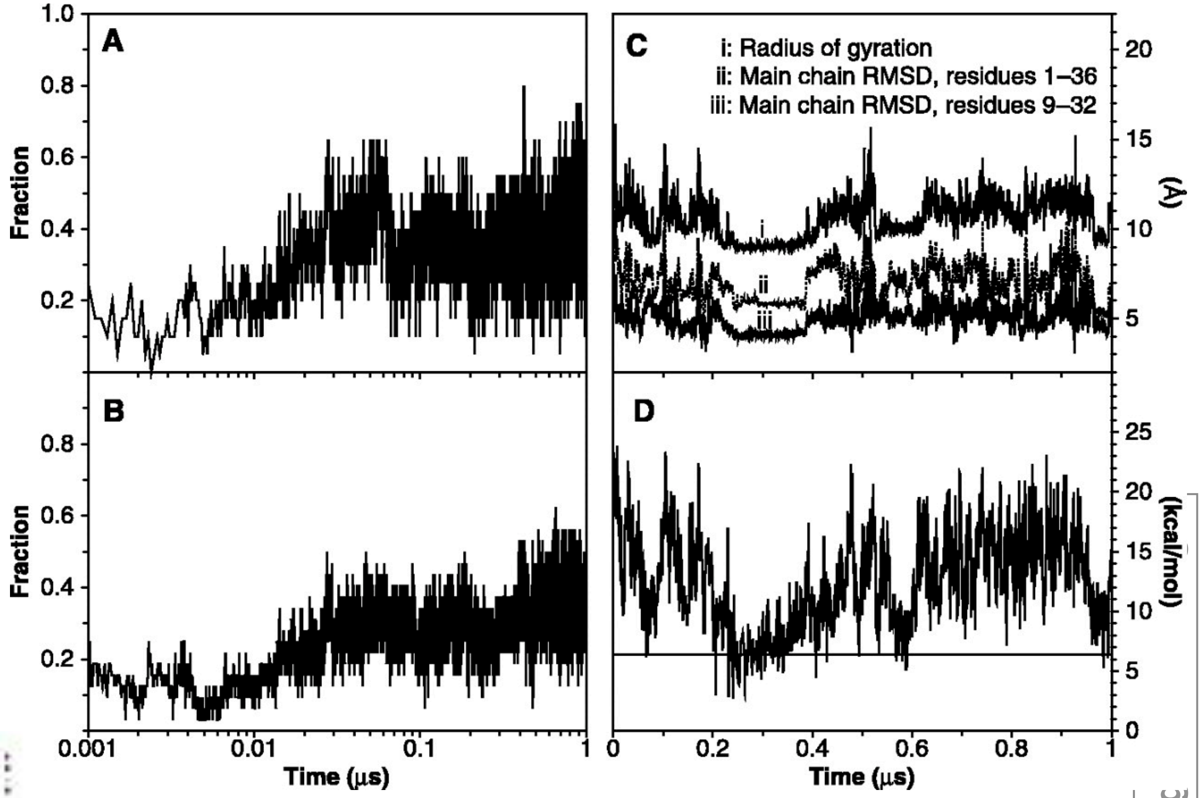
$$RMS(t) = \sqrt{\frac{\sum d_i^2(t)}{N}}$$

$$d_i(t) = \mathbf{R}(\mathbf{x}_i(t) - \mathbf{T}) - \mathbf{x}_i(0)$$

  - ◊ di = Difference in position of protein atom at t from the initial position, after structures have been optimally rotated translated to minimize RMS(t)
  - ◊ Solution of optimal rotation has been solved a number of ways (Kabsch, SVD)

Other Things to Calculate

- Fraction of Native Contacts
- Percent Helix
- Radius of Gyration



Caption: Time evolution of (A) fractional native helical content, (B) fractional native contacts, (C) R and the main chain rmsd from the native structure, and (D) SFE of the protein. The helical content and the native contacts are plotted on a logarithmic time scale. The helical content was measured by the main chain - angle
(60° ± 30°, 40° ± 30°). The native contacts were measured as the number of neighboring residues present in 80% of the last 50 ns of the native simulation. Residues are taken to be in contact if any of the atom pairs are closer than 2.8 Å, excluding residues i and i+1, which always have the contacts through main chain atoms. The SFE was calculated as described by Eisenberg and McLachlan (31) using their parameters (0.0163, 0.00637, 0.02114, 0.02376, and 0.05041, in kcal mol Å2, for the surface areas of nonpolar, polar, sulfur, charged oxygen, and charged nitrogen, respectively). The straight line represents the SFE of the native structure.

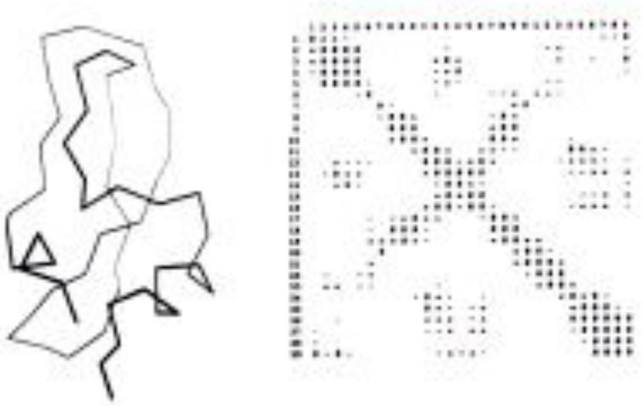


Illustration and Caption from Duan & Kollman (1998)

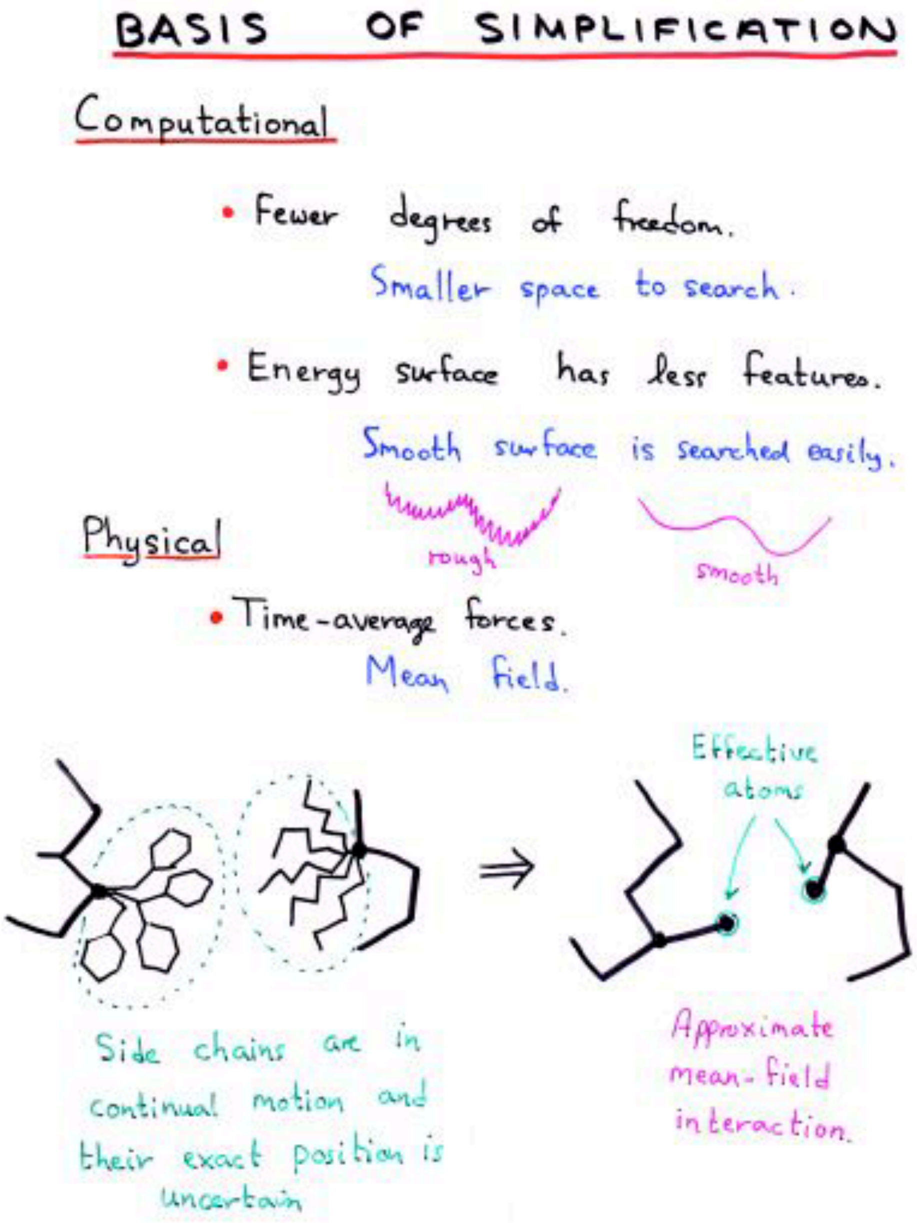
Timescales

Motion	length (Å)	time (fs)
bond vibration	0.1	10
water hindered rotation	0.5	1000
surface sidechain rotation	5	10^5
water diffusive motion	4	10^5
buried sidechain libration	0.5	10^5
hinge bending of chain	3	10^6
buried sidechain rotation	5	10^{13}
allosteric transition	3	10^{13}
local denaturation	7	10^{14}

Values from McCammon & Harvey (1987) and Eisenberg & Kauzmann

Simplified Simulation

Simplification



BASIS OF SIMPLIFICATION

Computational

- Fewer degrees of freedom.
  Smaller space to search.
- Energy surface has less features.
  Smooth surface is searched easily.

Physical

- Time-average forces.
  Mean field.

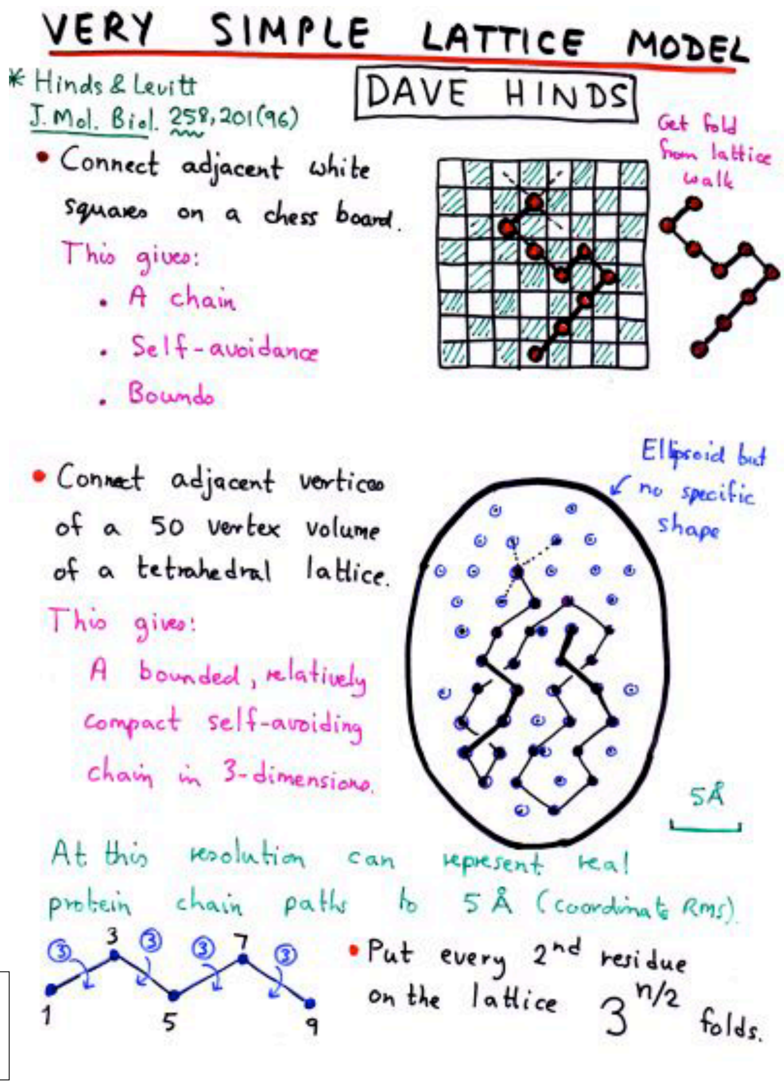
Side chains are in continual motion and their exact position is uncertain

Approximate mean-field interaction.

Illustration from M Levitt, Stanford University

Simplified Protein: Lattice Models

- Cubic Lattice
- Tetra-hedral Lattice



VERY SIMPLE LATTICE MODEL

DAVE HINDS

* Hinds & Levitt
J. Mol. Biol. 258, 201 (96)

- Connect adjacent white squares on a chess board.

This gives:

- A chain
- Self-avoidance
- Bounds

Get fold from lattice walk

- Connect adjacent vertices of a 50 vertex volume of a tetrahedral lattice.

This gives:

A bounded, relatively compact self-avoiding chain in 3-dimensions.

Ellipsoid but no specific shape

5Å

At this resolution can represent real protein chain paths to 5 Å (coordinate Rms).

- Put every 2nd residue on the lattice $3^{n/2}$ folds.

Illustration from M Levitt, Stanford University

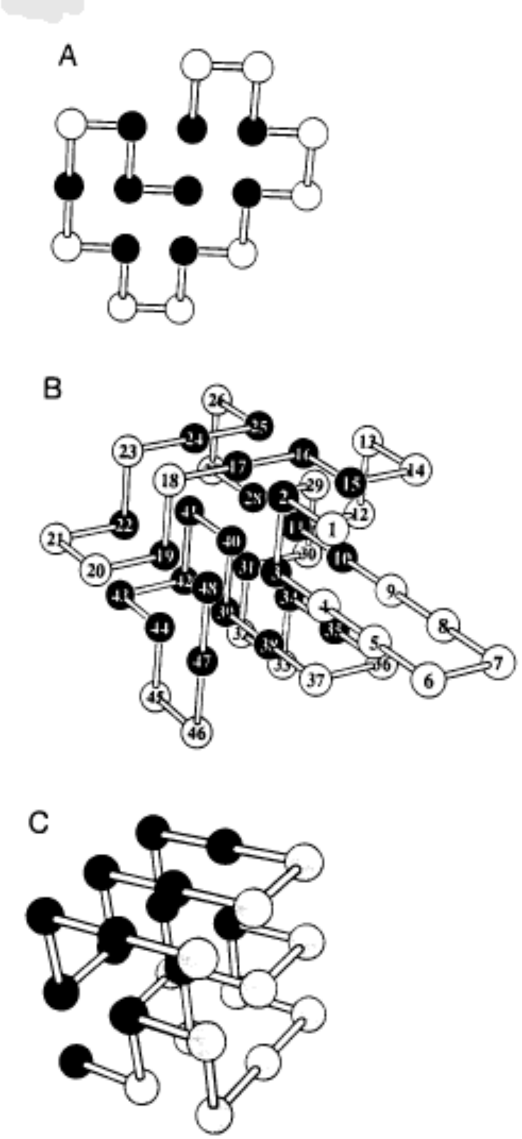
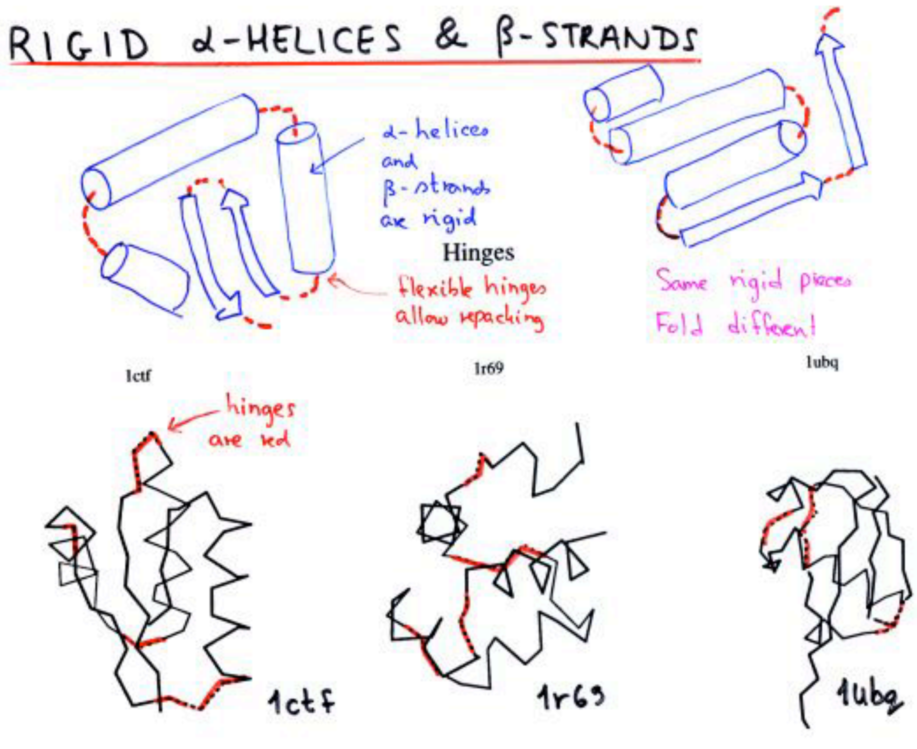


Illustration from Dill et al. (1990)

Off-lattice Discrete State Models



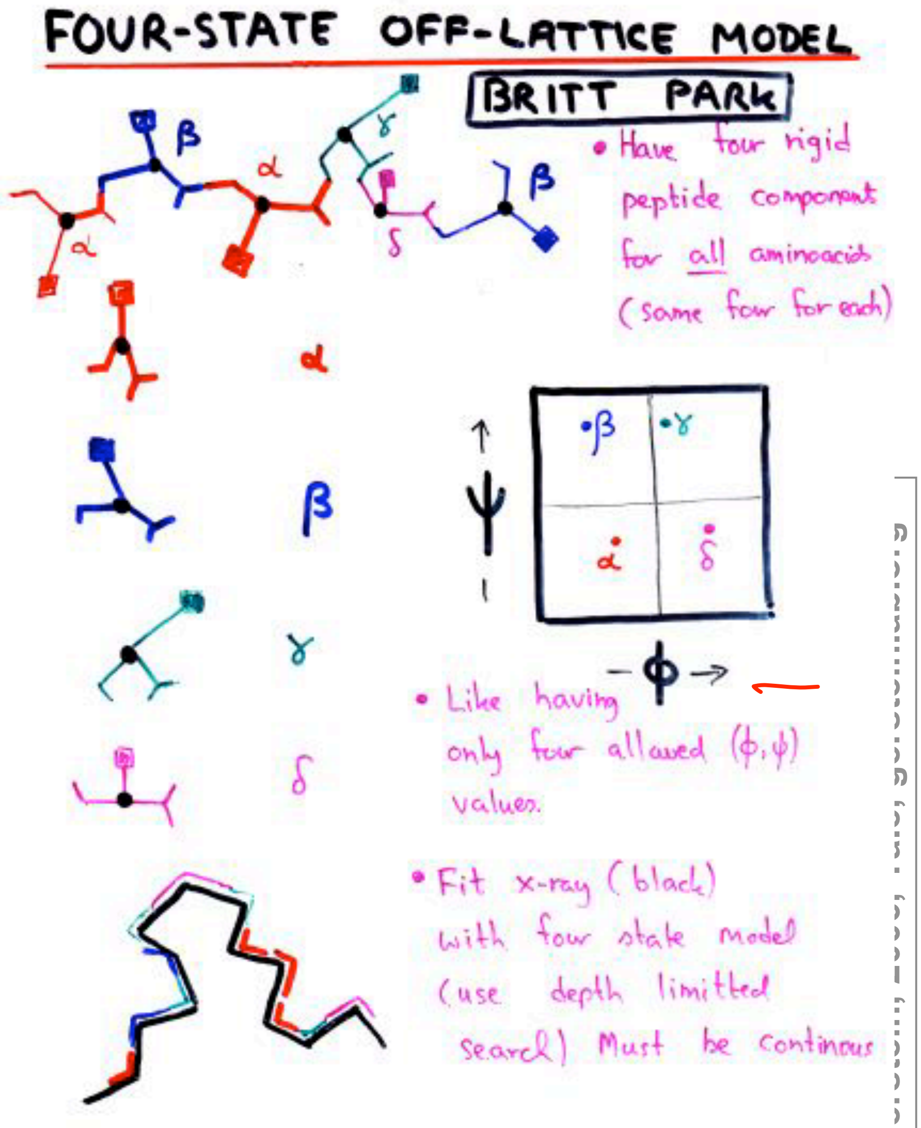
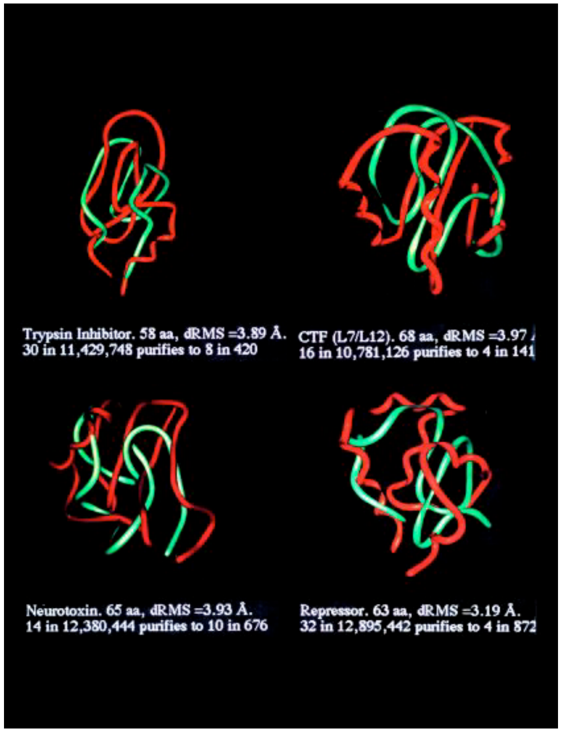


Illustration from M Levitt, Stanford University

How Well Do Lattice Structures Match Real Protein Structure?



Trypsin Inhibitor. 58 aa, dRMS =3.89 Å.
30 in 11,429,748 purifies to 8 in 420

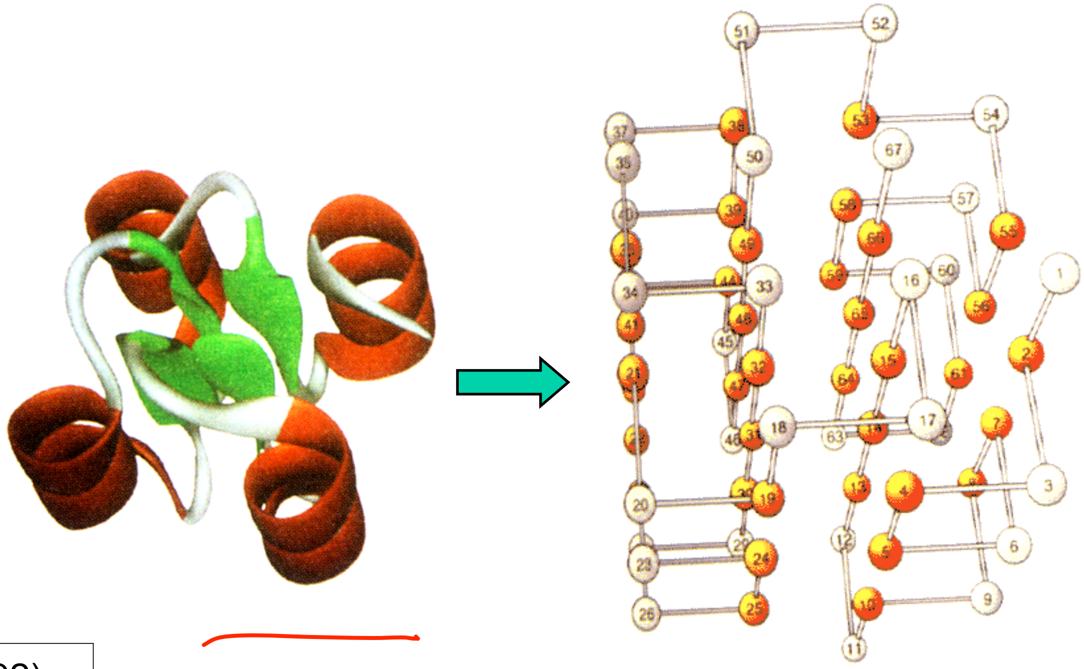
CTF (L7/L12). 68 aa, dRMS =3.97
16 in 10,781,126 purifies to 4 in 141

Neurotoxin. 65 aa, dRMS =3.93 Å.
14 in 12,380,444 purifies to 10 in 676

Repressor. 63 aa, dRMS =3.19 Å.
32 in 12,895,442 purifies to 4 in 872

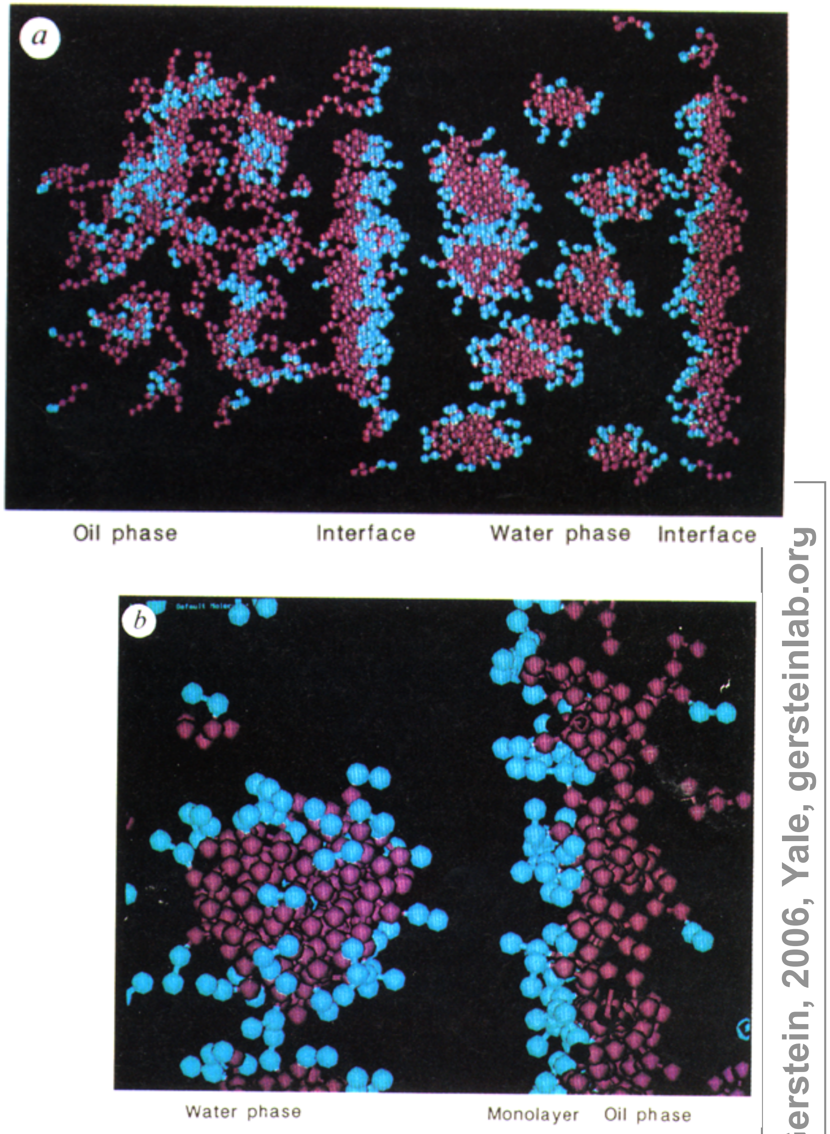
Illustration Credit: Hinds & Levitt (1992)

Illustration Credit: Dill et al. (1995)



Simplified Solvent

- Smit et al. (1990) Surfactant simulation
- Three types of particles, o, w and s
  - s consists of w-w-o-o-o-o
  - s has additional springs
- all particles interact through L-J potential
  - o-w interaction truncated so purely repulsive
- Above sufficient to give rise to the formation of micelles, membranes, &c



Figures from Smit et al. (1990)

End of class M9
[2006,11.29]
Start of class M11
[2006,12.06]

Electrostatics Revisited: the Poisson-Boltzmann Equation

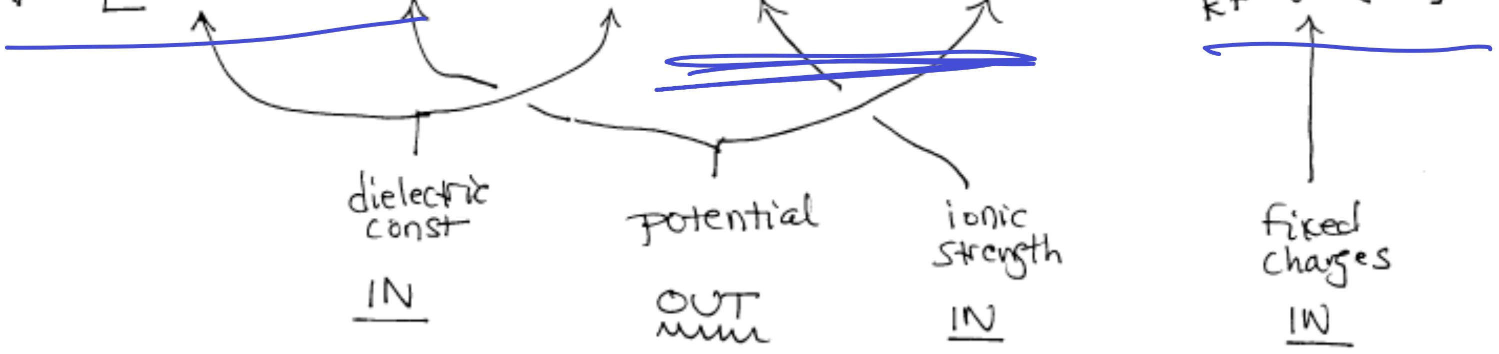
Poisson-Boltzmann equation

- Macroscopic dielectric
  - ◊ As opposed to microscopic one as for realistic waters
- Linearized: sinh $\phi = \phi$
  - ◊ counter-ion condense

- The model
  - ◊ Protein is point charges embedded in a low dielectric.
  - ◊ Boundary at accessible surface
  - ◊ Discontinuous change to a new dielectric (no dipoles, no smoothly varying dielectric)

PBE Eq. Ughh!

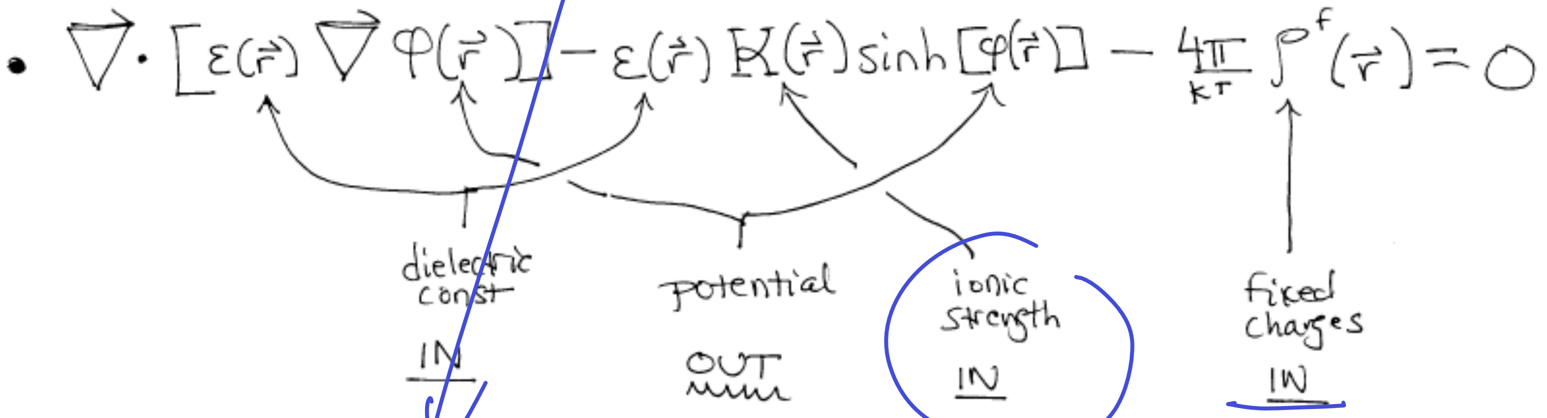
- $\nabla \cdot [\varepsilon(\vec{r}) \nabla \varphi(\vec{r})] - \varepsilon(\vec{r}) \bar{\kappa}(\vec{r}) \sinh[\varphi(\vec{r})] - \frac{4\pi}{kT} \rho^{f}(\vec{r}) = 0$



Simplifications of the Poisson-Boltzmann equation

- Laplace eq.
  - ◊ div grad V = ρ
  - ◊ grad V = **E** field
  - ◊ Only have divergence when have charge source

PBE Eq. Ughh!



- $\vec{\nabla}\cdot[\varepsilon(\vec{r})\vec{\nabla}\varphi(\vec{r})] - \varepsilon(\vec{r})\bar{\kappa}(\vec{r})\sinh[\varphi(\vec{r})] - \frac{4\pi}{kT}\rho^{f}(\vec{r}) = 0$

- No moving ions, Constant Dielectric → Poisson's Eq.

$$\boxed{\vec{\nabla}^2\varphi(\vec{r}) = \frac{4\pi}{kT\varepsilon}\rho(\vec{r})}$$

- Finite Difference Soln. to PDE

  (PDE has deriv. WRT to 2 var.
  ODE like Newton's Eq. has deriv. WRT to 1 var.)

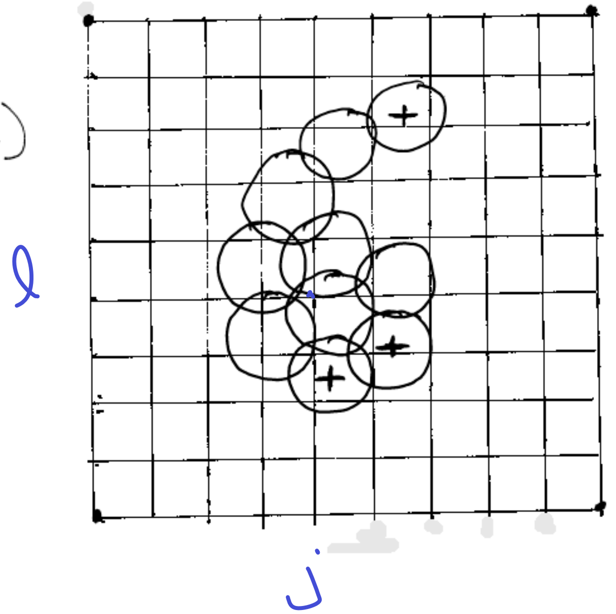
- $\vec{\nabla}^2 \varphi(\vec{r}) = \frac{4\pi}{kT\varepsilon}\rho(\vec{r})$

- in 2D $\frac{\partial \varphi}{\partial x^2} + \frac{\partial \varphi}{\partial y^2} = \overset{\text{const.}}{C}\rho(\vec{r})$

- $\frac{\partial^2 \varphi}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial \varphi}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{V_{j+1} - V_j}{\Delta}\right) = \frac{(V_{j+1} - V_j) - (V_j - V_{j-1})}{\Delta^2}$

  $(\partial x = \Delta,\ \partial \varphi = V_{j+1} - V_j)$

- $V_{j+1,l} + V_{j-1,l} + V_{j,l+1} + V_{j,l-1} - 4V_{j,l} = \Delta^2 C Q_{j,l}$



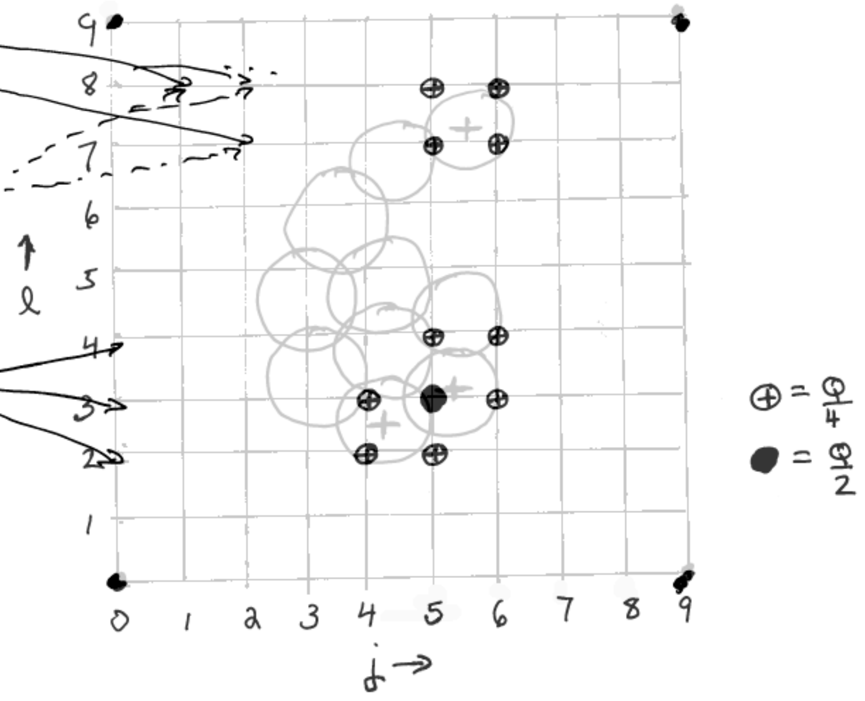
Protein on a Grid

For intuition ONLY -- Don't need to know in detail!!

9×9=81
$V_{j,l}$
unknown

9×9=81
$Q_{i,j}$
known

10+10+8+8=36
$V_{j,l}$
KNOWN
Boundary Conditions



Demand Consistency on the Grid

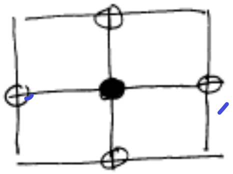
- $V_{j+1,l} + V_{j-1,l} + V_{j,l+1} + V_{j,l-1} - 4V_{j,l} = \Delta^2 C Q_{j,l}$
- System of Equations → solve for unknown $V_{j,l}$
- Matrix Inversion in Finite Diff. method

For intuition ONLY -- Don't need to know in detail!!

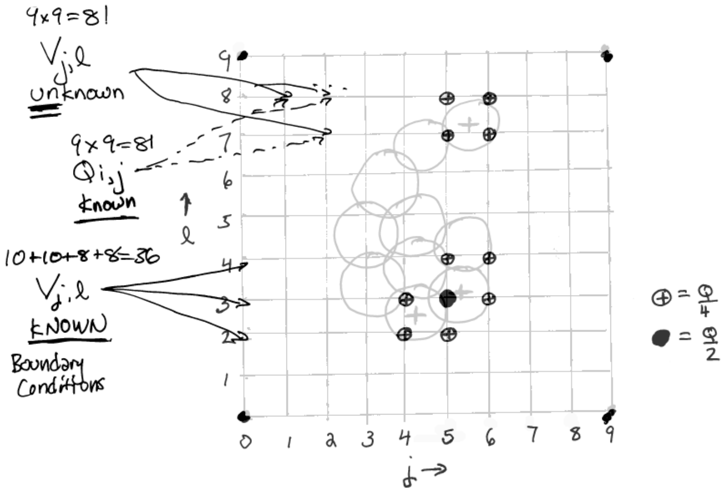
Relaxation: Deviation from consistency should vanish at $t \to \infty$

$$\nabla^2 V - 4\pi\rho = \boxed{\frac{\delta V}{\delta t}} \rightarrow 0 \text{ at } t=\infty$$

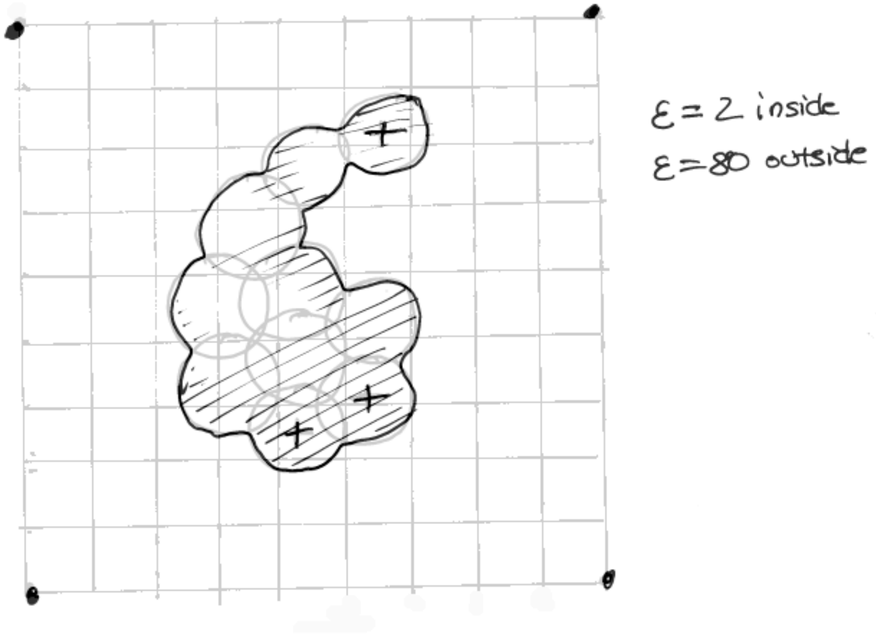
$$V_{j,l}^{t+1} \leftarrow V_{j,l}^{t} + \Delta t\left(\frac{V_{j+1,l}^{t} + V_{j-1,l}^{t} + V_{j,l+1}^{t} + V_{j,l-1}^{t} - 4V_{j,l}^{t}}{\Delta^2} - Q_{j,l}\right)$$

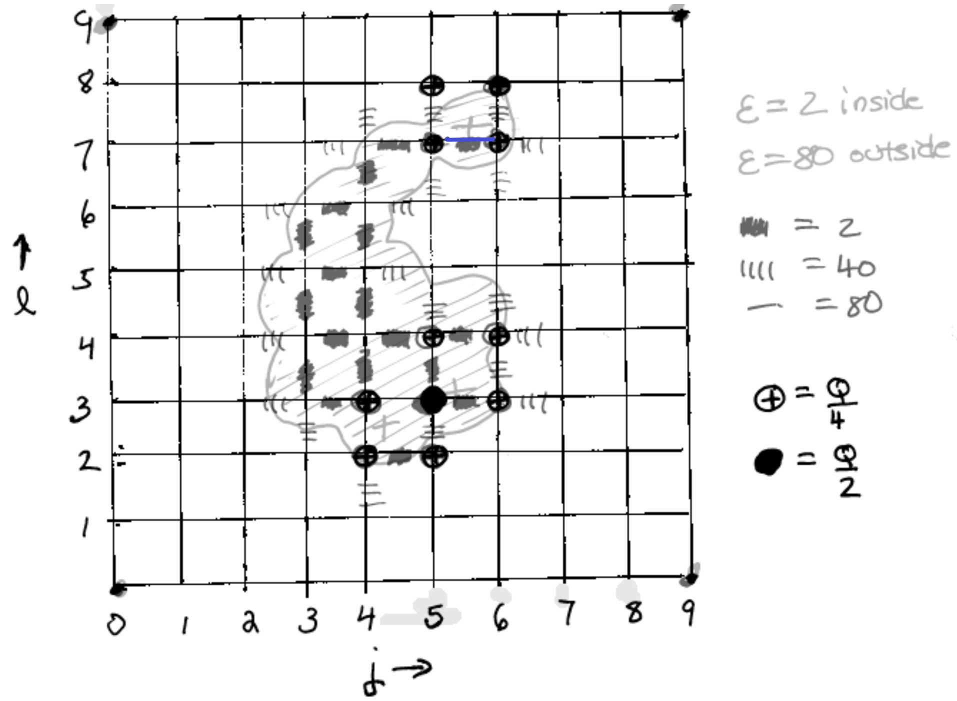


Avg value at center (●) is avg. value at 4 outside nodes (⊕) plus charge at center



Adding a Dielectric Boundary into the Model



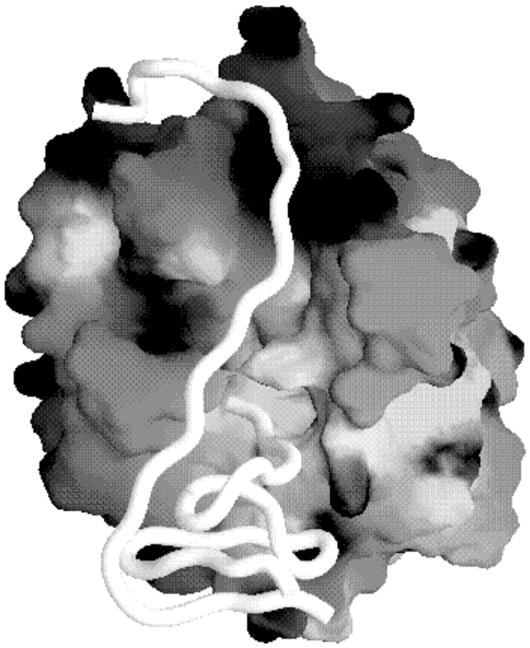


$$\nabla\cdot\left(\epsilon(r)\nabla\varphi\right) \Rightarrow$$
$$\frac{1}{\Delta^2}\Big(\epsilon(j\rightarrow j+1)(V_{j+1}-V_j) - \epsilon(j-1\rightarrow j)(V_j-V_{j-1})\Big)$$

Electrostatic Potential of Thrombin

The proteolytic enzyme Thrombin (dark backbone worm) complexed with an inhibitor, hirudin (light backbone worm). The negatively charged (Light gray) and positively charged (dark gray) sidechains of thrombin are shown in bond representation.

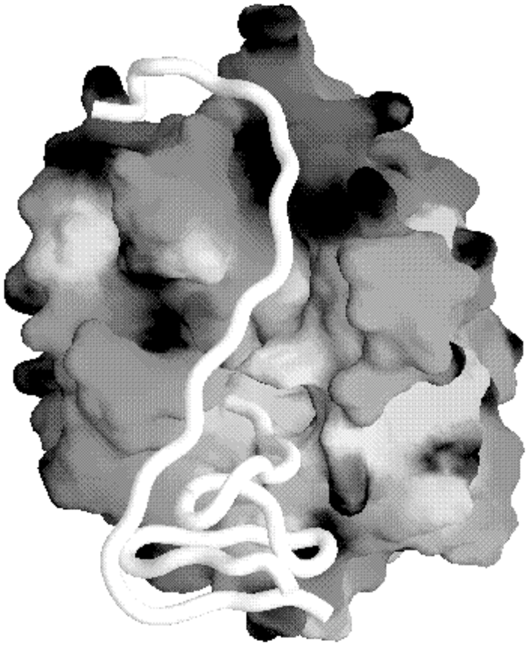
Graphical analysis of electrostatic potential distributions often reveals features about the structure that complement analysis of the atomic coordinates. For example, LEFT shows the distribution of charged residues in the binding site of the proteolytic enzyme thrombin. RIGHT shows the resulting electrostatic potential distribution on the protein surface. The basic (positive) region in the fibrinogen binding, while it could be inferred from close inspection of the distribution of charged residues in TOP, is more apparent in the potential distribution.



Solvent accessible surface of thrombin coded by electrostatic potential (dark: positive, light: negative). Hirudin is shown as a light backbone worm. Potential is calculated at zero ionic strength.

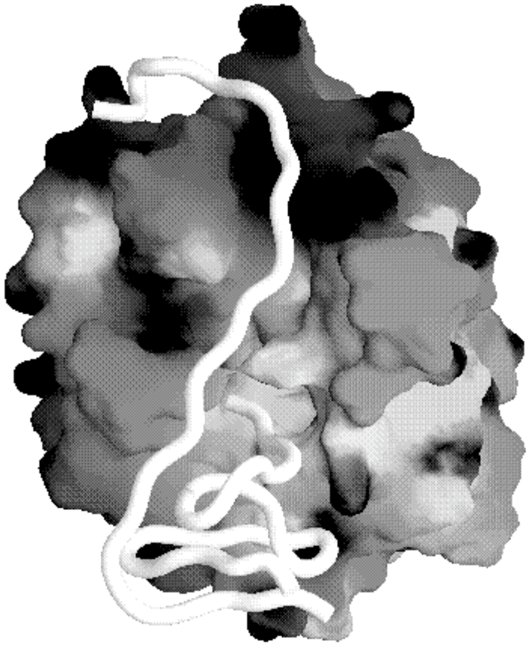
Illustration Credit: Sharp (1999)
Text captions also from Sharp (1999)

Increasing Ionic Strength



Solvent accessible surface of thrombin coded by electrostatic potential (dark: positive, light: negative). Hirudin is shown as a light backbone worm. Potential is calculated at physiological ionic strength (0.145M)

TOP shows the effect of increasing ionic strength on the potential distribution, shrinking the regions of strong potential in comparison to BOTTOM.

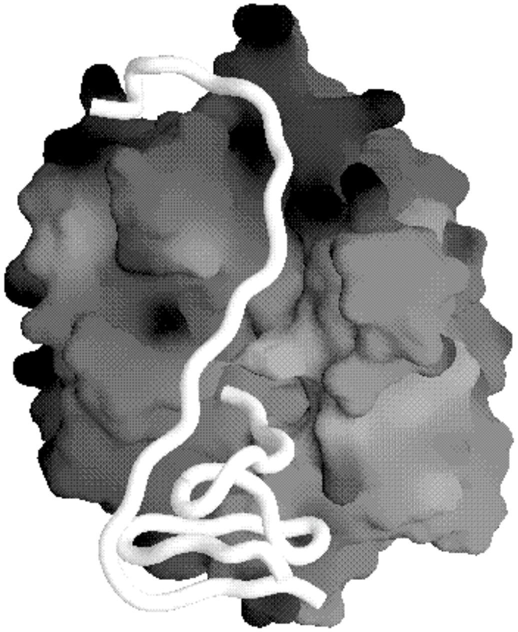


Solvent accessible surface of thrombin coded by electrostatic potential (dark: positive, light: negative). Hirudin is shown as a light backbone worm. Potential is calculated at zero ionic strength.

Illustration Credit: Sharp (1999)
Text captions also from Sharp (1999)

Increasing Dielectric

Solvent accessible surface of thrombin coded by electrostatic potential (dark: positive, light: negative). Hirudin is shown as a light backbone worm. Potential is calculated using the same polarizability for protein and solvent.



TOP is calculated assuming the same dielectric for the solvent and protein. The more uniform potential distribution compared to BOTTOM shows the focusing effect that the low dielectric interior has on the field emanating from charges in active sites and other cleft regions.

Solvent accessible surface of thrombin coded by electrostatic potential (dark: positive, light: negative). Hirudin is shown as a light backbone worm. Potential is calculated at zero ionic strength.

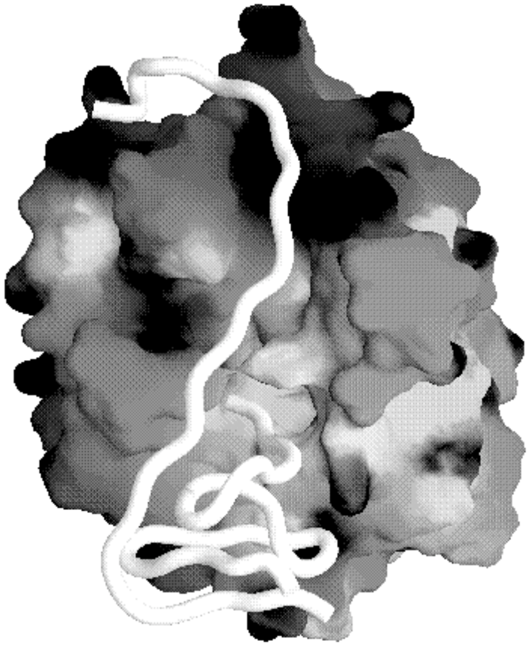
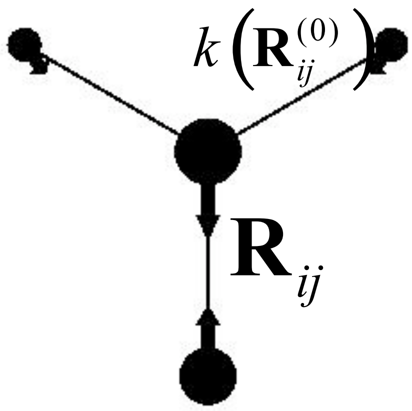


Illustration Credit: Sharp (1999)
Text captions also from Sharp (1999)

Normal Modes

NMA formalism and implementation:



$$U\left(\mathbf{R}_1 \ldots \mathbf{R}_N\right) = \sum_{i<j}^{N} U_{ij}\left(\mathbf{R}_i - \mathbf{R}_j\right)$$

$$U_{ij}(r) = k\left(\mathbf{R}_{ij}^{(0)}\right)\left[r - \left|\mathbf{R}_{ij}^{(0)}\right|\right]^2$$

Hinsen (1998):

$$k(r) = c \cdot \exp\left(-\frac{|r|^2}{r_o^2}\right)$$

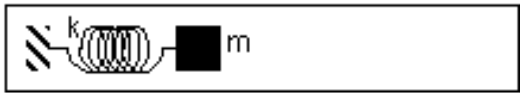
$$Solve: \quad \ddot{\mathbf{Q}}\,{+}\mathbf{A}\mathbf{Q} \quad = 0$$

Minimize pot. energy
Then diagonalize the 2nd derivative of the potential energy
Simplified potential

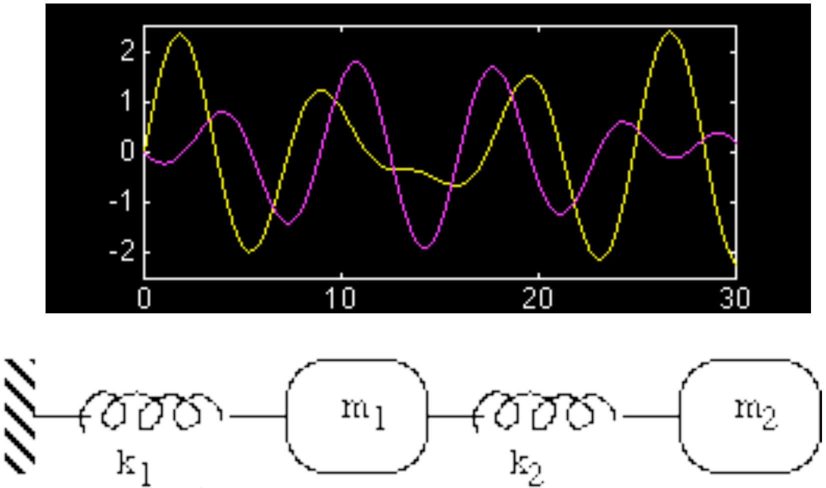
Normal Mode Analysis

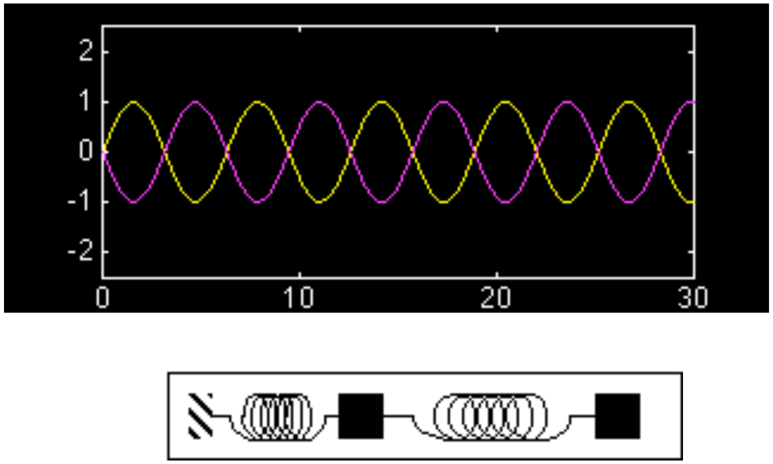
- Examine vibrational motions
  - ◊ For a system of N particles:
    - Total number of displacement = 3N
    - Total number of vibrational modes = 3N-6
- Mode frequency indicates the type of motion:
  - ◊ Low frequency: Collective motion
  - ◊ High frequency: Localized motion

What is “NMA”?

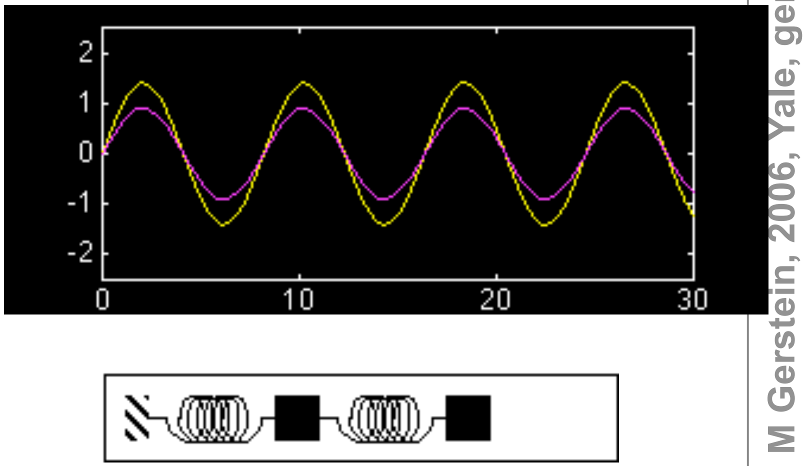


$$v = A\sin\left(\sqrt{\frac{k}{m}}t\right)$$



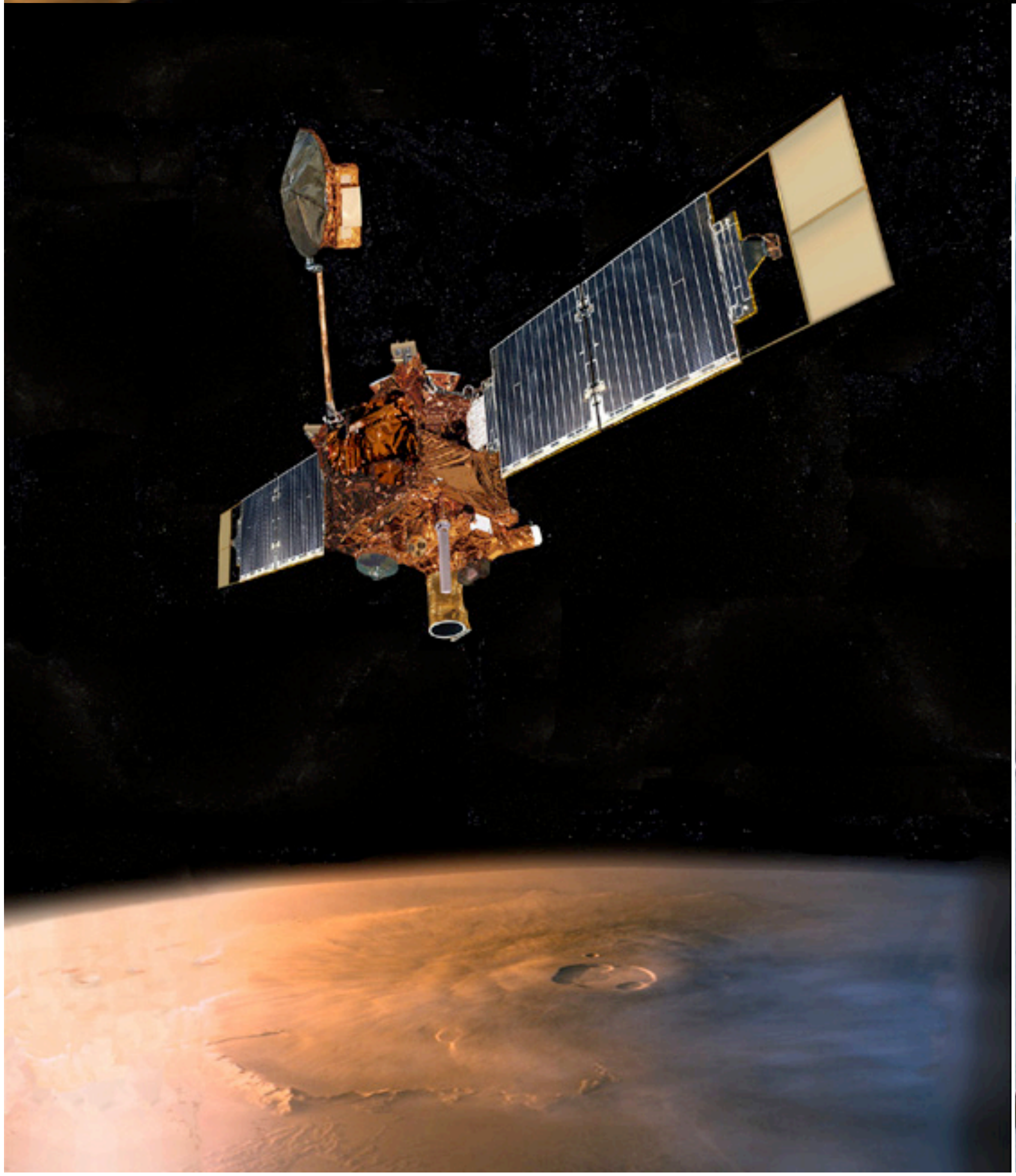


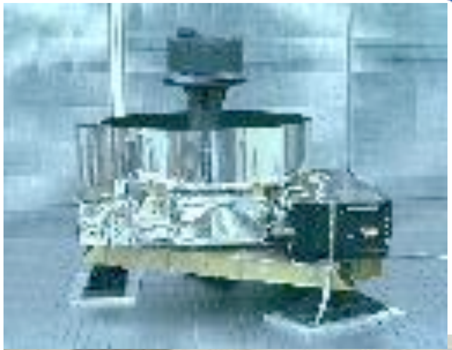
+

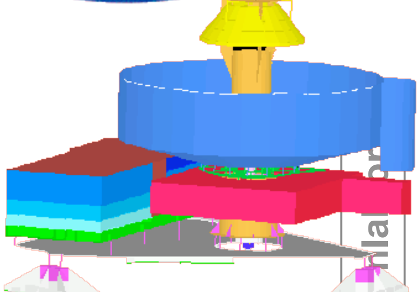


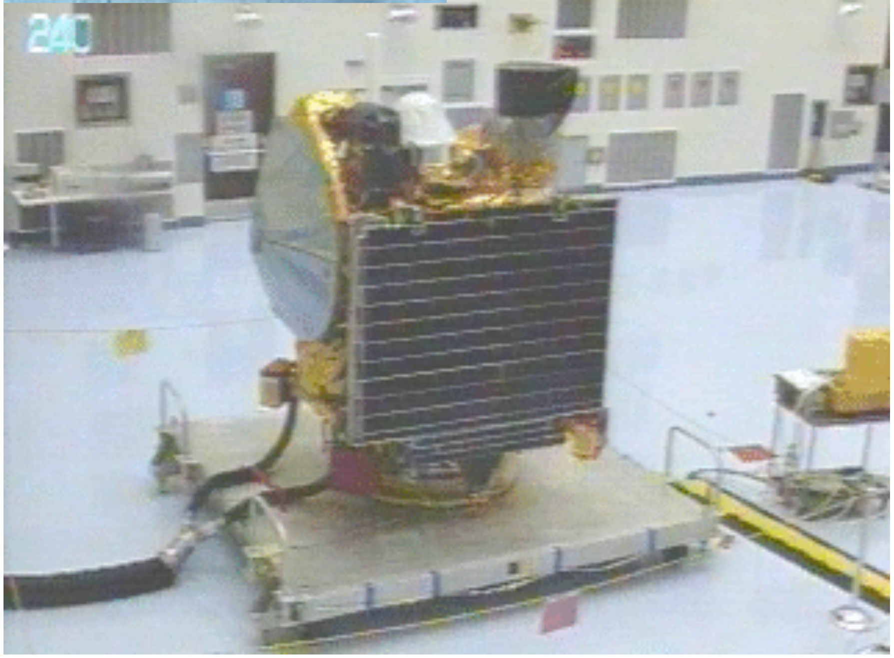
M O L A

Mars Orbiter Laser Altimeter • NASA/GSFC









Normal (*Natural*) Modes for MOLA