BIOCH 755: Biochemistry I Fall 2015

# 6. Molecular Mechanics and Molecular Dynamics Simulation

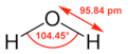
Jianhan Chen
Office Hour: M 1:30-2:30PM, Chalmers 034
Email: jianhanc@ksu.edu
Office: 785-2518

## Determinant of Structure (or Lack of It)

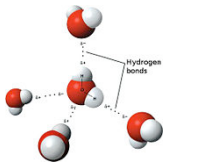
- Probability of observing a particular structure (conformation) is determined by its stability (as defined by the free energy)
  - Thermodynamics and statistical mechanics!
- No single structure is *the* structure
  - It is all about probability (statistical mechanics!)
  - Motions and flexibility are important too
- The stability depends on a range of factors
  - Intramolecular interactions
    - Bonded: chemical bonds, angles, dihedrals etc
    - Nonbonded: "weak" interactions
      - Charged-charged, van der Waals (dispersion and repulsion)
  - Intermolecular interactions: nonbonded/weak interactions
    - Cellular environment: solvent (water), membrane, salt, pH etc
    - Association with other biomolecules, small molecules, ions, etc

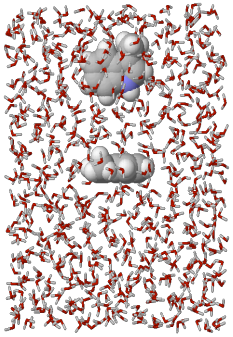


## Water




- Solvent of life
- Many unique properties
  - Maximum density at 4 °C
    - Ice is lighter than liquid water
  - Polar molecule
    - hydrogen bonding network
    - High specific heat capacity
- Hydrophobicity and hydrophilicity
  - Solute polarity (carry partial charges or not)
  - Salts (e.g. NaCl) dissolve in water readily
  - Hydrocarbons (oil) do not mix with water
- Amphipathic molecules
  - Self-assembly to micelles, biological membranes



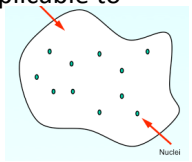





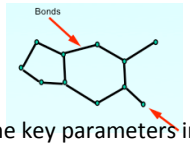
# MOLECULAR MECHANICS

# Quantum Mechanics vs. Molecular Mechanics

- Quantum mechanics: "exact" and most applicable to understand chemical reactions
  - Separate nuclei and electrons
  - Too expensive, and not sufficiently accurate
  - Not relevant as many biological processes




- Molecular mechanics: classical mechanics at molecular level
  - Classical treatment of all atoms
  - No electron, no chemistry
  - Allows description of large molecules
  - Experimental methods available to determine the key parameters in a molecular mechanical treatment

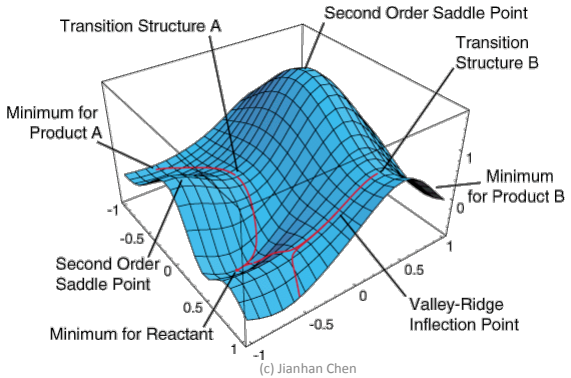



- Hybrid QM/MM
  - QM for the active site (where reaction occurs) and MM for the rest
  - Accurate treatment of MM/QM Boundary is a problem



# Classical Mechanics

- Total energy: $E = K + V$
  - Kinetic energy ($K = mv^2/2$), potential energy V (i.e., force field)
- Newton's second law of motion: $F = m\,a$
  - Relation of force and potential energy: $F = -\delta V/\delta r$



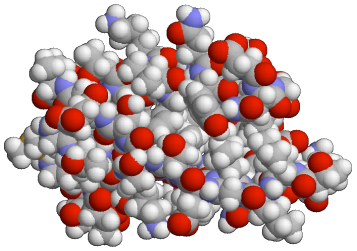


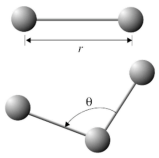

# Molecular Potentials

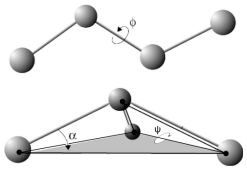
- Basic form: $V = V_{bonding} + V_{nonbonding}$

  $= ( \Sigma V_{bond} + \Sigma V_{angle} + \Sigma V_{dihe}) + \Sigma(V_{elec} + V_{vdw})$

  - The potential energy is a function of all coordinates.
  - Additivity, empirical, transferability



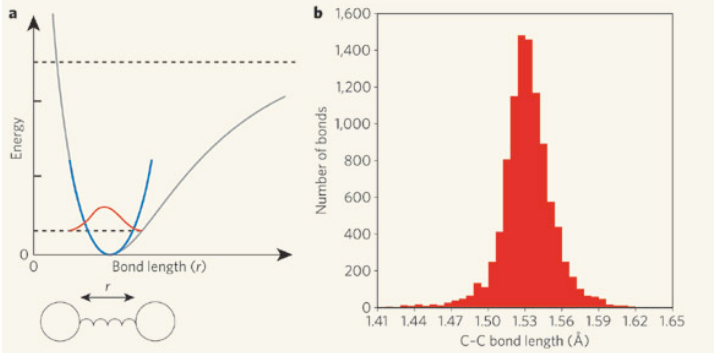




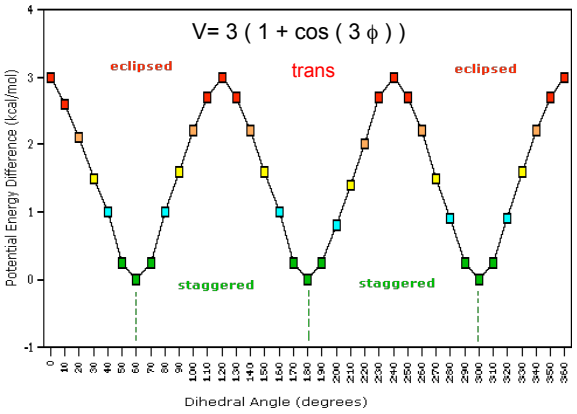

# Bonds and Angles

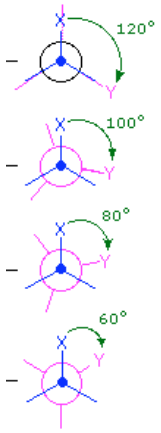
- $V_{bond} = k_{bond} ( r - r_o)^2$
  - Harmonic approximation
  - OK for biomolecules
- $V_{angle} = k_{angle} ( \theta - \theta_o)^2$



# Dihedral Potentials

- $V_{dihe} = k_{dihe} \cdot [1 + \cos(n\phi - \delta)]$

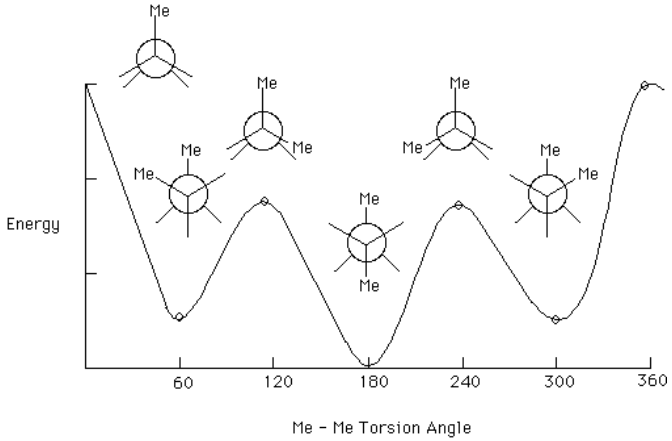





# Realistic Dihedral Potentials

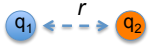
- Actual dihedral potentials often have contributions with multiple periodicities



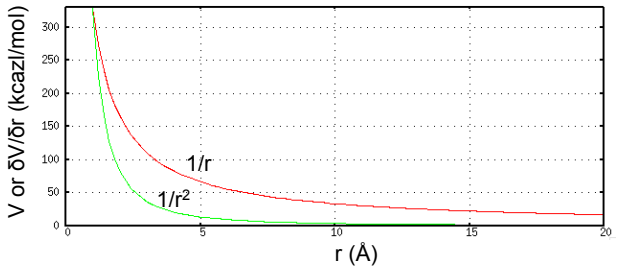



# Electrostatic Interactions

- $V_{elec} = q_1q_2/4\pi\varepsilon_0 r$ Coulomb's Law
  - $\varepsilon_0$: permittivity constant of vacuum



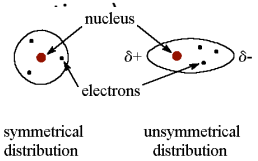

- A simplified form: $V_{elec} = 332\ q_1q_2/r$
  - Where *q* is unit of electron charge, *r* is in Å, and *V* in kcal/mol.
- Dielectric medium: $V_{elec} = 332\ q_1q_2/\varepsilon r$
  - ε is dielectric constant (relative permittivity).
  - ε=78 for water under lab conditions (300K, 1atm)






# van der Waals Interactions

- London dispersion: attractive forces that arise from temporary dipoles (induced dipole-induced dipole int



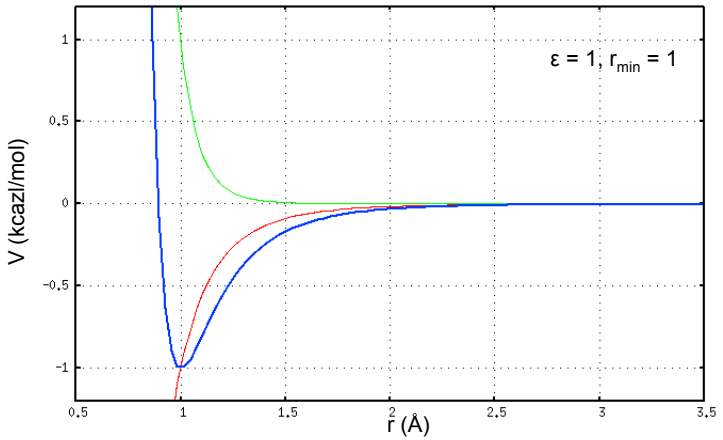


- van der Waals repulsion: all atoms repel at short distances
- A common function form: $V_{vdw} = -A/r^6 + B/r^{12}$
- Lennard-Jones potential function (12-6)

$$V(r) = \epsilon \left[ \left( \frac{r_{min}}{r} \right)^{12} - 2 \left( \frac{r_{min}}{r} \right)^{6} \right]$$

## Lennard-Jones Potential

$$V(r) = \epsilon \left[ \left( \frac{r_{min}}{r} \right)^{12} - 2 \left( \frac{r_{min}}{r} \right)^{6} \right]$$






## Sub-Summary

- Molecular mechanics as an effective approach for calculating the energies of large biomolecules
- Common functional form with five terms
  - Covalent bonded terms
  - Non-bonded terms: electrostatics, vdW
- Coming up: application of molecular mechanics in biomolecular modeling and simulation (i.e., molecular dynamics)

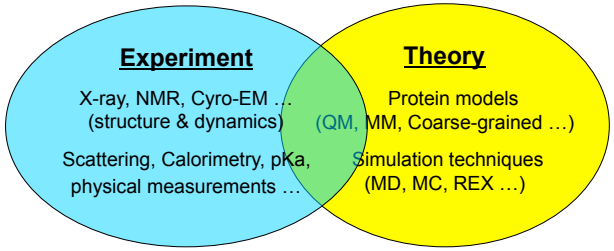


# MOLECULAR DYNAMICS



## Why Modeling?

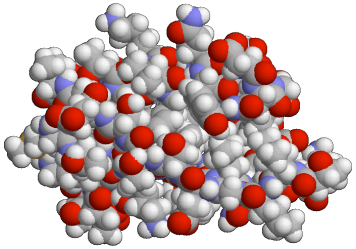
- Visualization
- Interpretation of experimental data
  - X-Ray Crystallography, NMR Structures, EM models
  - Low resolution experimental data: mutagenesis, FRET, and many more
- Novel insights not accessible to experiments

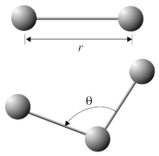


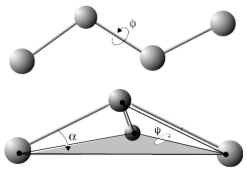

- To rationalize and predict experimental observations.
- To provide new insights, hypothesis, and directions.

# Molecular Potentials

- Basic form: $V = V_{bonding} + V_{nonbonding}$
  $= ( \Sigma V_{bond} + \Sigma V_{angle} + \Sigma V_{dihe}) + \Sigma(V_{elec} + V_{vdw})$
  - The potential energy is a function of all coordinates.
  - Additivity, empirical, transferability





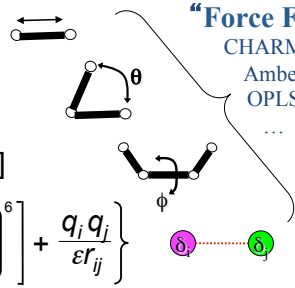






# Molecular Mechanics

## Classical Energy Functions

$$V_{MM} = \sum_{bonds,i} \frac{1}{2} k_i^b \cdot (b_i - b_i^0)^2$$
$$+ \sum_{angles,i} \frac{1}{2} k_i^\theta \cdot (\theta_i - \theta_i^0)^2$$
$$+ \sum_{torsions,i} k_i^\phi \cdot [1 + \cos(n_i \phi_i - \delta_i)]$$
$$+ \sum_{atoms,i} \sum_{i<j} \left\{ \varepsilon_{min}^{ij} \left[ \left( \frac{r_{min}^{ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{min}^{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon r_{ij}} \right\}$$




"Force Field"
CHARMM
Amber
OPLS
…

**Molecular Dynamics (MD)**

$$m_i \ddot{r}_i = F_i = -\nabla V_i$$

**Monte Carlo (MC)**

$$P(\delta r) = \exp(-\Delta V/kT)$$



# CHARMM param22 Force Field

- Topology file: define the building blocks (atoms, connectivities)

atom types

residue blocks

atom compositions

connectivity

```
MASS     1 H      1.00800 ! polar H
MASS     2 HC     1.00800 ! N-ter H
MASS     3 HA     1.00800 ! nonpolar H
…
                                  name  type  charge
RESI ALA              0.00
GROUP
ATOM N    NH1    -0.47  !     |
ATOM HN   H       0.31  !  HN-N
ATOM CA   CT1     0.07  !     |     HB1
ATOM HA   HB      0.09  !     |    /
GROUP                   !  HA-CA--CB-HB2
ATOM CB   CT3    -0.27  !     |    \
ATOM HB1  HA      0.09  !     |     HB3
ATOM HB2  HA      0.09  !   O=C
ATOM HB3  HA      0.09  !     |
GROUP                   !
ATOM C    C       0.51
ATOM O    O      -0.51
BOND CB CA  N  HN  N  CA  O  C
BOND C  CA  C  +N  CA HA  CB HB1  CB HB2  CB HB3
IMPR N -C CA HN  C CA +N O
DONOR HN N
…
```

excerpted from: top_all22_prot.inp



# CHARMM param22 Force Field

- Parameter file: define the parameters of interactions

```
…
BONDS
C    C     600.000     1.3350 ! ALLOW ARO HEM
                ! Heme vinyl substituent (KK, from propene (JCS))
CA   CA    305.000     1.3750 ! ALLOW   ARO
                ! benzene, JES 8/25/89
…
ANGLES
CA   CA   CA    40.000    120.00   35.00   2.41620 ! ALLOW   ARO
                ! JES 8/25/89
CE1  CE1  CT3    48.00    123.50   !
               ! for 2-butene, yin/adm jr., 12/95
…
DIHEDRALS
C    CT1  NH1  C        0.2000  1   180.00 ! ALLOW PEP
                ! ala dipeptide update for new C VDW Rmin, adm jr., 3/3/93c
C    CT2  NH1  C        0.2000  1   180.00 ! ALLOW PEP
                ! ala dipeptide update for new C VDW Rmin, adm jr., 3/3/93c
…
NONBONDED nbxmod  5 atom cdiel shift vatom vdistance vswitch -
cutnb 13.0 ctofnb 12.0 ctonnb 10.0 eps 1.0 e14fac 1.0 wmin 1.5
                !adm jr., 5/08/91, suggested cutoff scheme
C      0.000000  -0.110000     2.000000 ! ALLOW   PEP POL ARO
                ! NMA pure solvent, adm jr., 3/3/93
CA     0.000000  -0.070000     1.992400 ! ALLOW   ARO
                ! benzene (JES)
```

excerpted from: par_all22_prot.inp

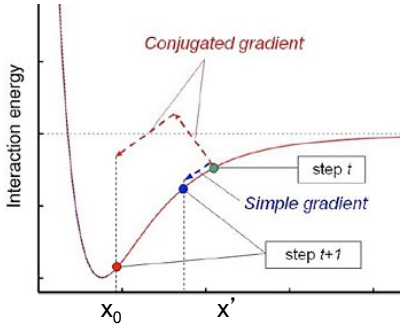
# Parameterization of Force Fields

- Bonded terms: spectroscopy or quantum mechanics
- Lennard-Jones: Small molecular crystals
- Electrostatic: quantum mechanics (fit monopoles to electrostatic potential)
- Many challenges in practice
  - which (model) molecules: availability, representative or not
  - how many atomic classes: transferability and tractability
  - Which properties to parameterize for?
  - Correlation of parameters
  - Higher order/new terms or not?
  - Electron polarization, non-additivity, …
  - water, water, and water
- At the end, do they add up? (cancellation of errors)



# Energy Minimization

- Minimization follows gradient of potential to identify stable points on energy surface
  - Let $V(x) = k (x-x_0)^2$
  - Begin at $x'$, how do we find $x_0$ if we don't know V(x) in detail?
    - How can we move from $x'$ to $x_0$?
  - Steepest descent (SD):
    - $x' \rightarrow x' = x+\delta$
    - $\delta = -dx\, \partial V(x)/\partial x = -dx\, k(x- x_0)$
  - This moves us, depending on the step size dx, toward $x_0$.
  - On a simple harmonic surface, we will reach the minimum, $x_0$, i.e. converge, in a certain number of steps related to dx.






# Molecular Dynamics

- Objective: $\{r_1(t), \ldots, r_N(t)\} \rightarrow \{r_1(t+\Delta t), \ldots, r_N(t+\Delta t)\}$ f = ma
- Basic idea: solve Newton's equation of motion numerically
  - Given current coordinates (x), velocities (v)
    - Forces can be calculated based on coordinates (from $f = -\partial V/\partial x$)
    - $x(t+\Delta t) = x(t) + v(t)\,\Delta t$
    - $v(t+\Delta t) = v(t) + f(t)/m\,\Delta t$
    - Repeat above operations
- More accurate integrators (better energy conservation)
  - Verlet Algorithm (Verlet J. Chem. Phys. 1967)

consider Taylor's expansions:

$x(t\pm\Delta t) = x(t) \pm v(t)\,\Delta t + 1/2m\, f(t)\,\Delta t^2 \pm 1/6\, d^3x/dt^3\,\Delta t^3 + O(\Delta t^4)$

Adding expansion $x(t+\Delta t)$ and $x(t-\Delta t)$ and rearrange:

$x(t+\Delta t) = 2x(t) - x(t-\Delta t) + f(t)/m\,\Delta t^2 + O(\Delta t^4)$

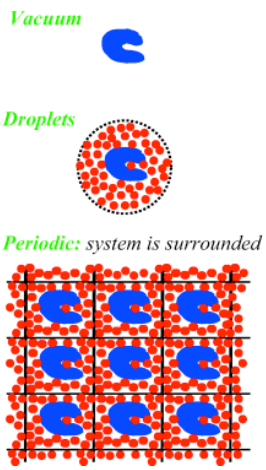
Subtracting expansion $x(t+\Delta t)$ and $x(t-\Delta t)$ and rearrange:

$v(t) = [x(t+\Delta t)-x(t-\Delta t)/(2\Delta t) + O(\Delta t^3)$

velocities lag



# Solvent and Periodic Boundary Conditions




- Surface effects (surface tension)
- No dielectric screening

- Still surface effects (at water – vacuum interface)
- Only partial dielectric screening
- Evaporation of the solvent

Advantage:
- No surface effects

Disadvantage:
- Artificial periodicity
- High effective concentration

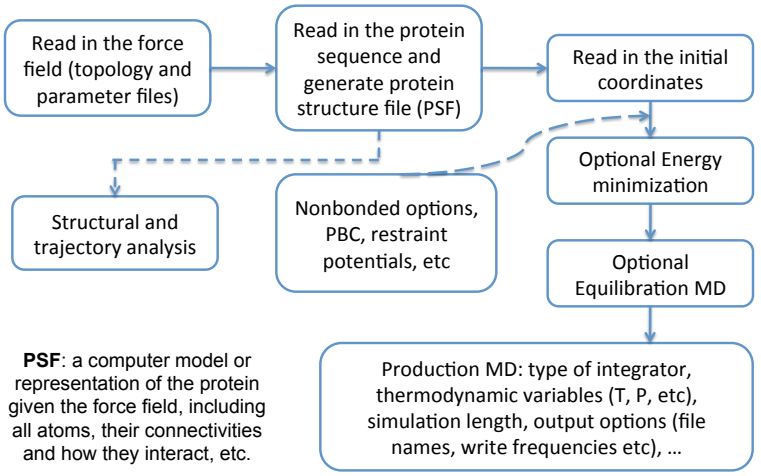
van Gunsteren Angew Chem Int Ed (2006)

## Controlling Thermodynamic Variables

- MD generate statistical ensembles that connect microscopic details to macroscopic/thermodynamic properties
- NVE (microcanonical - Entropy rules!)
- NVT (Canonical - Helmholtz free energy is relevant, A)
  - temperature $T = \sum m\langle v2\rangle/(3k_B)$
- NPT (Isothermal-isobaric - Gibbs free energy is relevant, G)
  - P=kinetic + virial contributions
- Thermostats, barostats, etc., allow one to choose appropriate ensembles
  - Following Nose', Hoover, Evans and others…
  - See Brooks, Curr. Opin. Struct. Biol., 5, 211(1995)]



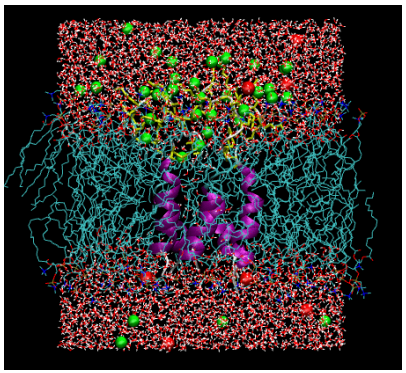
## Basic Flow of a MD Simulation




**PSF**: a computer model or representation of the protein given the force field, including all atoms, their connectivities and how they interact, etc.



## Biomolecular Simulations are computationally very expensive



Channel-forming peptides
in a fully solvated membrane bilayer;
Channel: 1795 atoms; All: 26254 atoms

**Simulated Time**

1 ns ($10^{-9}$ s)
(500,000 MD steps)

**CPU Time**

~200 hours ($10^6$ s)

**Wall Time**

~1 days ($10^5$ s) / 8 CPUs

- very small time step required
  - $\delta t$ ~ fs ($10^{-15}$ s)
- interactions between thousands of atoms need to be computed

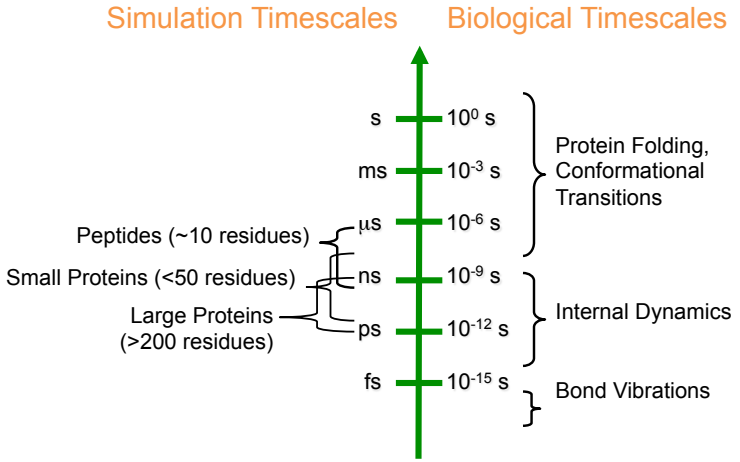


## Biological Time Scale

- Bond vibrations 1 fs ($10^{-15}$ s)
- Sugar repuckering 1 ps ($10^{-12}$ s)
- DNA bending 1 ns ($10^{-9}$ s)
- Domain movement 1 ms ($10^{-6}$ s)
- Base pair opening 1 ms ($10^{-3}$ s)
- Transcription 2.5 ms / nucleotide
- Protein synthesis 6.5 ms / amino acid
- Protein folding ~ 10 s (speed limit: μs)
- RNA lifetime ~ 300 s

Simulation time should exceed the time scale of interest by ~10-fold !

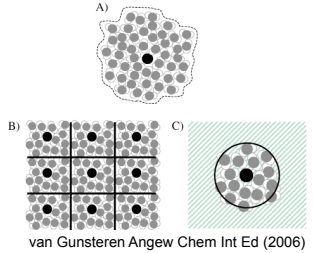
# Gap in Timescales



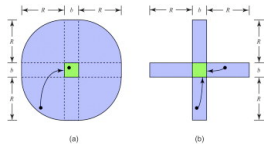



# Practical Considerations

- Long-range forces
  - Using cut-off to reduce the number of nonbonded atom pairs (~ 12-15Å)
  - Electrostatic decays slowing (1/r) and cut off does not work well; Particle Mesh Eward (PME) is needed.
- Parallel execution
  - Partition various regions of the system to different CPUs
  - Need to communicate information between nodes; this is a bottleneck
- Simplifications of the model
- Enhanced sampling techniques



van Gunsteren Angew Chem Int Ed (2006)


Shaw JCC (2005)

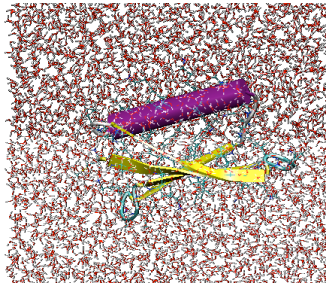


# Anton Computer

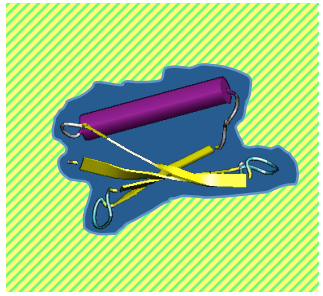
- Specialized computer for molecular dynamics simulations
- Highly parallel with extensive specialized hardware for MD related computations
- Rumored to cost >$100M to design and build
- 17 μs/day for small proteins (~25K atoms)
- ~5 μs/day for large proteins (500 aa, >132K atoms)
- A 512-node Anton donated by DESRES is available for public use at Pittsburg Supercomputer Center


# Implicit Solvent

- Solvent increases the system size about 10-fold
- It is possible to describe the mean influence of water w/o explicitly including water



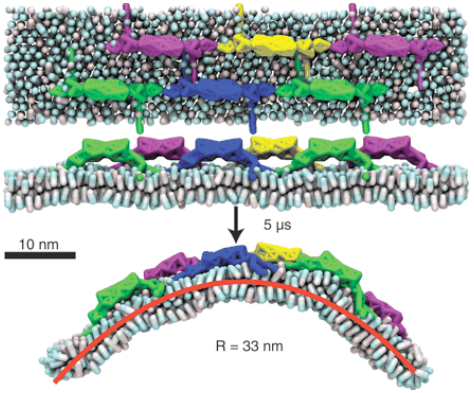
**Explicit solvent**
Protein: 56 residues (855 atoms)
Solvent: 5411 waters (16233 atoms)



**Implicit solvent**
Hybrid macroscopic (solvent) / microscopic (solute)

# Coarse-Grained Models

- Rely on reduced representation and/or simplified interaction schemes to access larger length and time scales



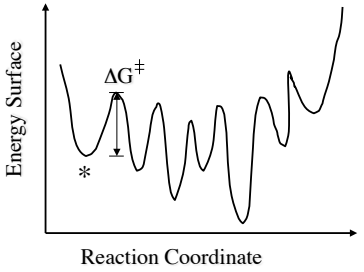

**Biomembrane sculpting by protein-BAR domains**

The simulation shown in the figure was carried out using a box with dimensions 100 x 16 x 50 nm and would correspond to a system of 10 million atoms. Using a shape-based CG model reduces the size to 3265 CG beads. The simulation showed that a concerted action of BAR domains arranged in a lattice results in the development of a global membrane curvature on a time scale of several μs, with the resulting curvature radius of ~30 nm that was observed experimentally.

Klein and Shinoda, Science (2008)



# Barriers, Temperature and Timescales




$$\tau = \tau_0 \exp(\Delta G^{\ddagger}/kT)$$

$\tau_0 \sim 10^{-12}$ s ~ ps

$\underline{T = 300\ K}$

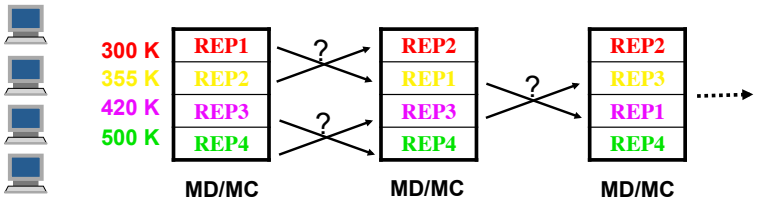
$\Delta G^{\ddagger}$: 1 kcal/mol, $\tau \sim$ ps
5 kcal/mol, $\tau \sim$ ns
10 kcal/mol, $\tau \sim$ μs

Protein energy landscape is highly complex and rugged with numerous local minima.



# Enhanced Sampling Techniques

**Replica Exchange (REX)**

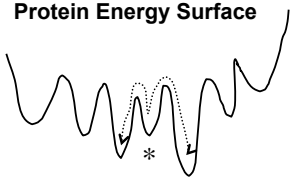



**Exchange criteria**

$$P_{i \Leftrightarrow j} = \begin{cases} 1 & \Delta \le 0 \\ \exp(-\Delta) & \Delta > 0 \end{cases}$$

$$\Delta = (E_i - E_j) \bullet (1/kT_j - 1/kT_i)$$

**Protein Energy Surface**



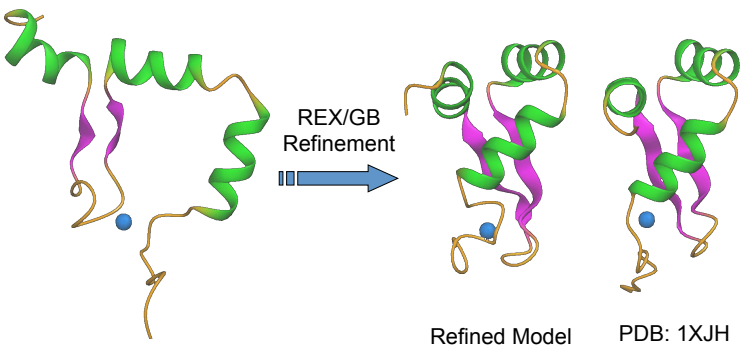

Sugita and Okamoto, *CPL* (1999); MMTSB Tool Set: *http://mmtsb.scripps.edu*



# Applications of Modeling

- Main advantages
  - Offer atomistic spatial resolution and femtosecond time resolution
  - Allow probing the system in many nontrivial ways that are not possible or too dangerous experimentally
  - Often much cheaper than doing the experiment itself
  - Can be applied at very large scales (computers are cheap)
  - Can provide theoretical frameworks for experimental studies
- A few prototypical areas
  - Protein structure prediction and calculation
  - Virtual screening and rational drug design
  - Simulation of important systems: mechanisms
  - Interpretation of (static) experimental data
  - Protein misfolding and aggregation
  - Biomolecular engineering: design of new enzymes etc
  - …

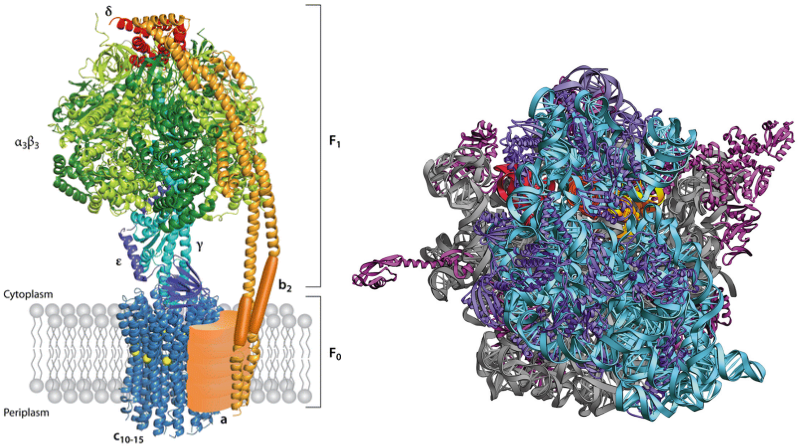
# NMR Structure Refinement




Chen et. al., *JACS* (2004); Chen et al., *J. Biomol NMR* (2004).



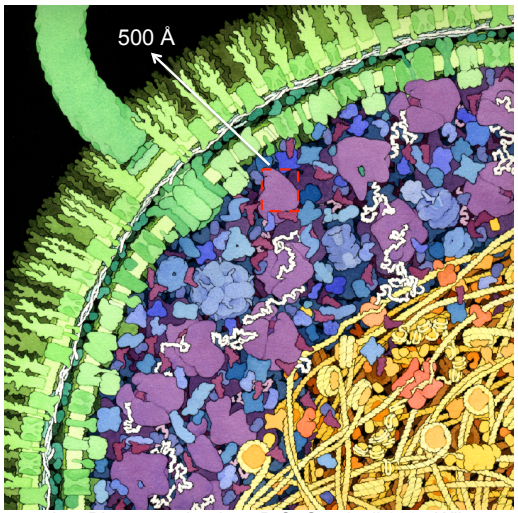
# Protein Folding Problem




von Ballmoos C, et al. 2009.
Annu. Rev. Biochem. 78:649–72



# Inconclusive Experiments




see also, Bionanotechnology D.S. Goodsell 2004 Wiley

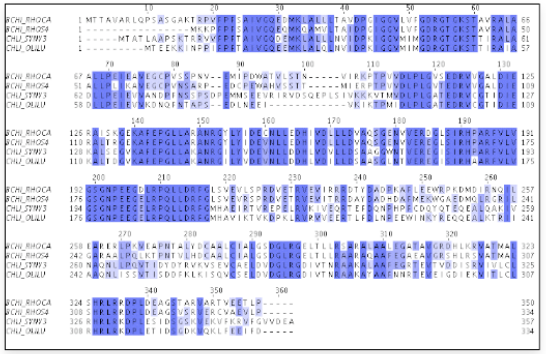
Extreme simplification

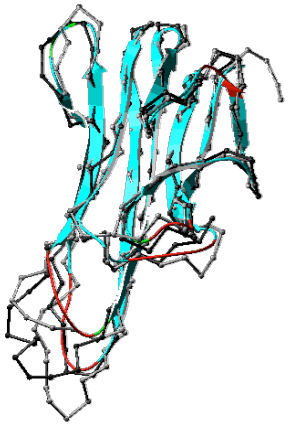
Limited force field accuracy

Large gaps in timescales



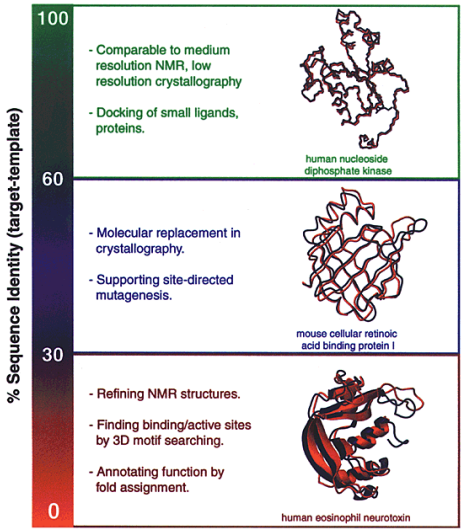
# Homology Modeling

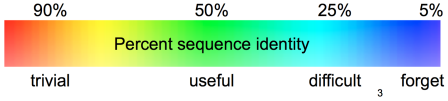




## Basic Principles

- Proteins with similar sequences should have similar structure.
- One can model a sequence of unknown structure (target) based on a homolog of known structure (template).
- Structure -> function

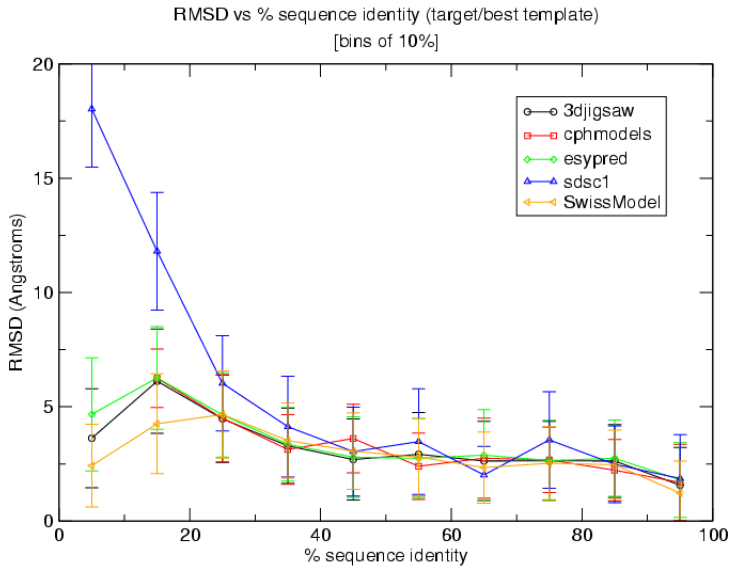



Sali, A. & Kuriyan, J. *Trends Biochem. Sci.* **22**, M20–M24 (1999)








http://swissmodel.expasy.org/?pid=smd03

## PDB is "complete" (i.e., novel fold is rare)

PNAS PNAS PNAS

### On the origin and highly likely completeness of single-domain protein structures

**Yang Zhang*, Isaac A. Hubner†, Adrian K. Arakaki*, Eugene Shakhnovich†, and Jeffrey Skolnick*‡**

*Center of Excellence in Bioinformatics, University at Buffalo, State University of New York, 901 Washington Street, Buffalo, NY 14203; and †Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Edited by Harold A. Scheraga, Cornell University, Ithaca, NY, and approved December 30, 2005 (received for review October 27, 2005)

**The size and origin of the protein fold universe is of fundamental and practical importance. Analyzing randomly generated, compact sticky homopolypeptide conformations constructed in generic simplified and all-atom protein models, all have similar folds in the library of solved structures, the Protein Data Bank, and conversely, all compact, single-domain protein structures in the Protein Data Bank have structural analogues in the compact model set. Thus, both sets are highly likely complete, with the protein fold universe arising from compact conformations of hydrogen-bonded, secondary structures. Because side chains are represented by their $C^{\beta}$ atoms, these results also suggest that the observed protein folds are insensitive to the details of side-chain packing. Sequence specificity enters both in fine-tuning the structure and thermodynamically stabilizing a given fold with respect to the set of alternatives. Scanning the models against a three-dimensional active-site library, close geometric matches are frequently found. Thus, the presence of active-site-like geometries also seems to be a consequence of the packing of compact, secondary structural elements. These results have significant implications for the evolution of protein structure and function.**

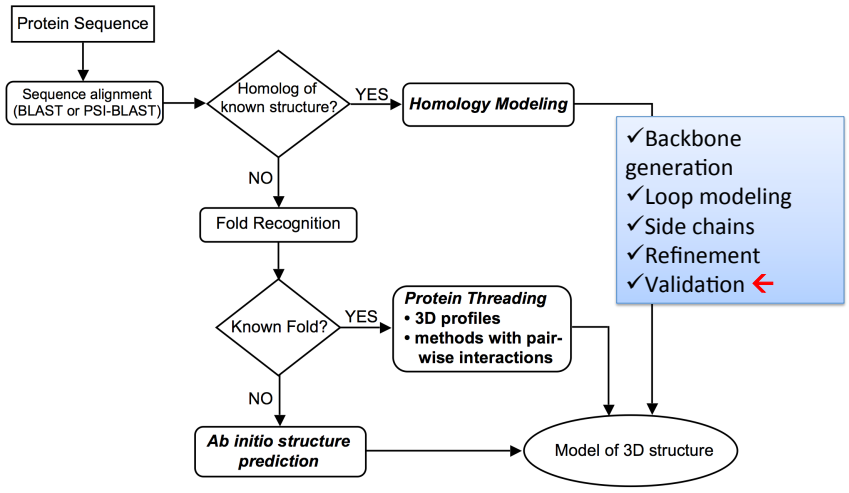
protein structure space is extremely dense in that there are many apparently nonhomologous structures that give acceptable structural alignments to an arbitrary selected single-domain protein. However, the structural alignment usually has unaligned regions or gaps. Starting from these alignments, state-of-the-art refinement algorithms can build full-length models that are of biological utility [with an average root-mean-square deviation (rmsd) to native of 2.3 Å for the backbone atoms] (14). Furthermore, incorrectly folded models generated by structure prediction algorithms also have structural analogues in the PDB, an observation again consistent with PDB completeness (15). Nevertheless, one might argue that comparing PDB structures against themselves as well as with structures generated using knowledge-based potentials extracted from the PDB (which retain some features of native proteins), although suggestive that the PDB is complete, does not establish that the universe of single-domain protein structures is complete; nor even if true, does it establish the reason for such completeness.

Here, we address these issues and show the surprising result that the highly likely completeness of the PDB results from the





## Homology Modeling Basic Flow Chart




Adapted in part from figure in http://www.cs.wright.edu/~mraymer/cs790/Homology_Modeling.ppt

## Structure Validation

- Covalent geometry: typically OK
- Ramanchandron plot: usually OK
- Inside/outside distributions of polar and apolar residues can be useful.
- Biological/biochemical data
  - Active site residues
  - Modification sites
  - Interaction sites
- Validation servers/tools:
  - ProQ
  - WhatIF
  - Procheck



## Popular Structure Prediction Servers

- Modern prediction tools/servers employ sophisticated integration of homology modeling, de novo modeling, structure refinement and many other empirical "tricks" to get the most out of existing statistical and physical knowledge
- Rosetta/Robetta (David Baker)
- I-TASSER (Yang Zhang)
- MULTICOM (Jianlin Cheng)