Modeling of Cryo-EM Maps

Workshop

Baylor College of Medicine

Klaus Schulten, U. Illinois at Urbana-Champaign

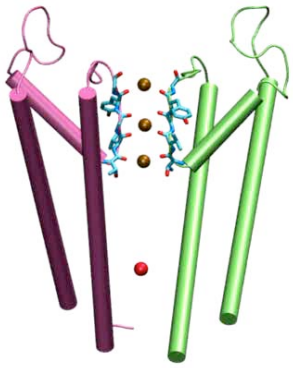
Molecular Modeling Flexible Fitting 1: Introduction to Molecular Dynamics

$$U(\vec{R}) = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} +$$
$$\underbrace{\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]}_{U_{dihedral}} +$$
$$\underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}}$$

MD force field

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

Equation of motion

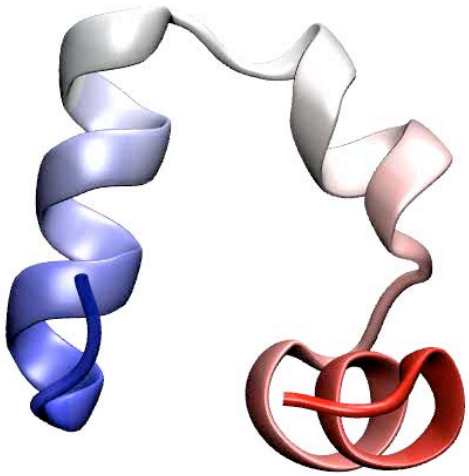


Simulated system

Phillips *et al.*, *J. Comp. Chem.* **26**:1781-1802, 2005.

Folding WT villin *in silico*

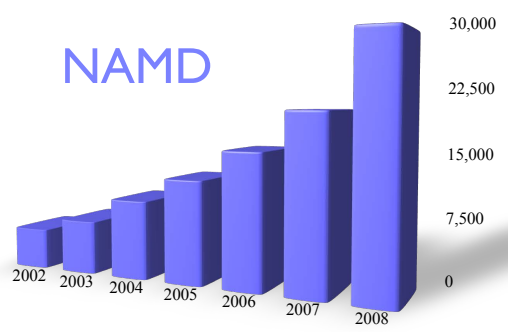
Three folding simulations reach native state within 5-8 μs



Peter L. Freddolino and Klaus Schulten. *Biophysical Journal*, 97:2338-2347, 2009.

Software Widely Used by Scientific Community

Sustained professional software development effort shipping products used by over 150,000 researchers/students worldwide



28,898 registered users
13,160 website visitors/month
1,200 citations

121,391 registered users
22,600 website visitors/month
3,000 citations

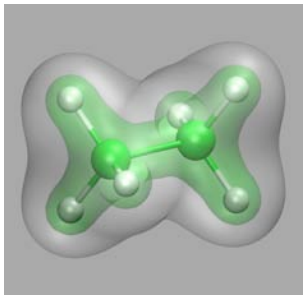
Team: K. Schulten (Physics), L. Kalé (Computer Sciences), Z. Schulten (Chemistry), R. Brunner, J. Phillips, J. Stone, K. Vandivort, D. Hardy, C. Harrison, B. Isralewitz, J. Saam, P. Freddolino, L. Trabuco

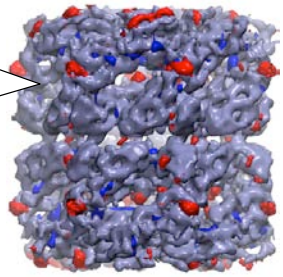
National Center for Research Resources

VMD – A Tool to Think

Volumetric Data:
Density maps,
Electron orbitals,
Electrostatic potential,
Time-averaged occupancy, …

23,000 Users

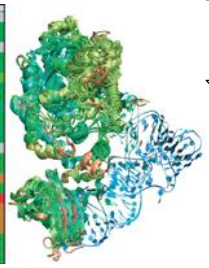




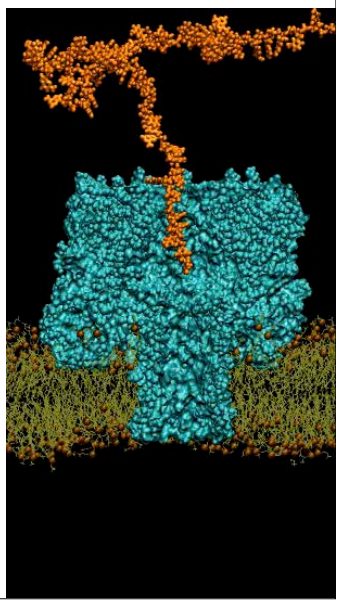
Sequence Data:
Multiple Alignments,
Phylogenetic Trees

VMD

Annotations



Atomic Data:
Coordinates,
Trajectories,
Energies,
Forces, …



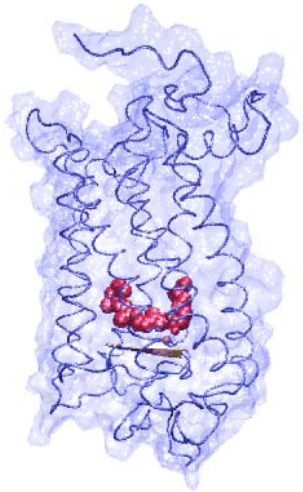
National Center for Research Resources

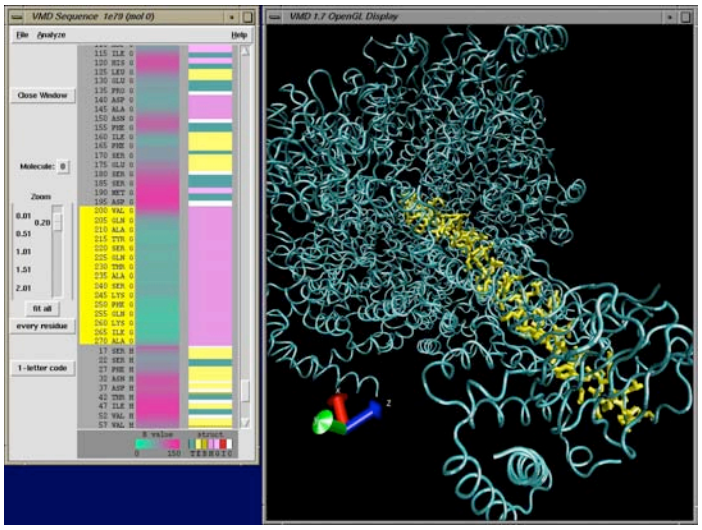
Key Features of VMD

- General 3-D molecular visualization with extensive drawing and coloring methods
- Extensive atom selection syntax for choosing subsets of atoms for display
- Visualization of dynamic molecular data
- Visualization of volumetric data
- Supports all major molecular data file formats
- No limits on the number of molecules or trajectory frames, except available memory
- Molecular analysis commands for structure, sequence, and dynamics
- Rendering high-resolution, publication-quality molecule images
- Movie making capability
- Building and preparing systems for molecular dynamics simulations
- Interactive molecular dynamics simulations
- Extensions to the Tcl/Python scripting languages
- Extensible source code written in C and C++

Molecular Graphics Perspective of Protein Structure and Function

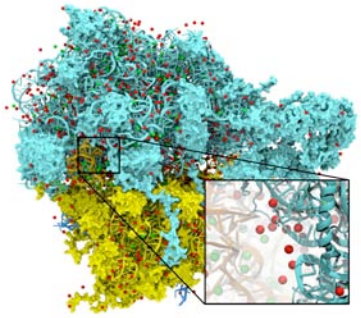
see tutorial at http://www.ks.uiuc.edu/Training/Tutorials/



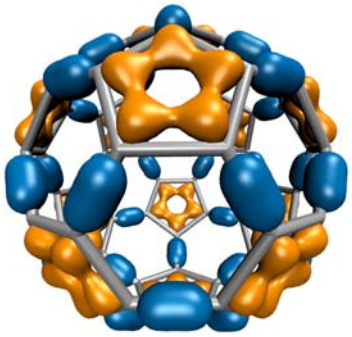


animation sequence structure

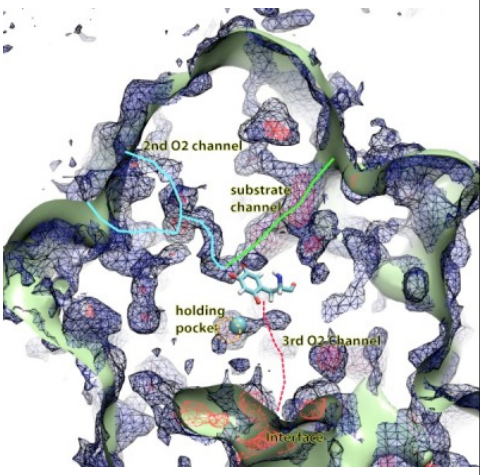
CUDA+OpenCL Acceleration in VMD



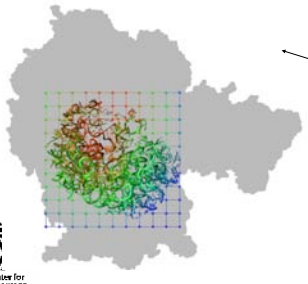
Electrostatic field calculation,
Multilevel Summation Method
20x to 44x faster

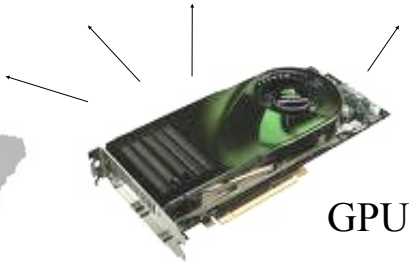


Molecular orbital
calculation and display
100x to 120x faster



Imaging of gas migration pathways in proteins with Implicit Ligand Sampling (ILS) algorithm
20x to 30x faster

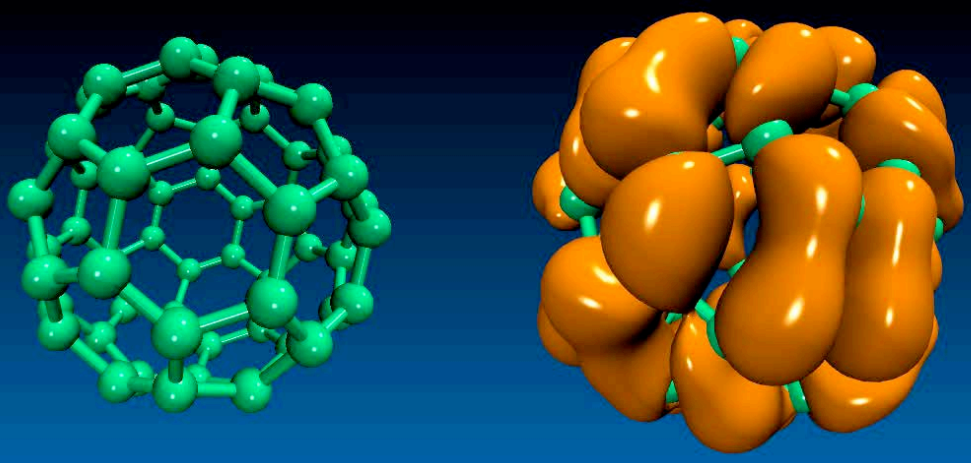




GPU

Computing and Visualizing Molecular Orbitals on GPUs

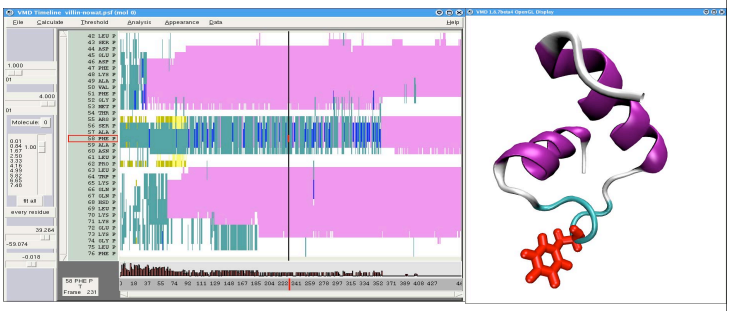
Quantum chemical calculation: Ivan Ufimtsev and Todd Martinez, Stanford U.



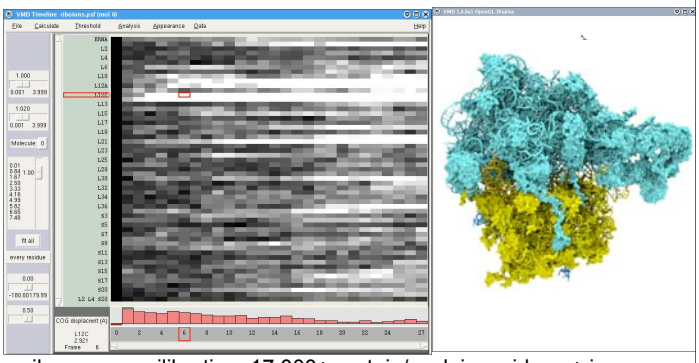
The future is here: Classical and quantum mechanical simulations

Analysis of Terascale/Petascale Simulation Data: VMD Timeline Tool

- Overview image shows varying properties over entire structure, trajectory
- Many analysis methods available; user-extensible
- Supports remote analysis; multi-terabyte trajectories don't need to be transferred from supercomputers
- Select areas of interest; only selected regions need to be transferred for analysis



7μs simulation; over 1 million frames to examine

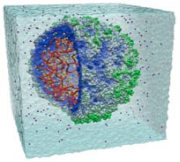


ribosome equilibration, 17,000+ protein/nucleic residues + ions

A Word on NAMD: Scalable Molecular Dynamics

Phillips *et al.*, *J. Comp. Chem.* **26**:1781-1802, 2005.

- Practical supercomputing for biomedical research
  - 32,000 users can't all be computer experts.
  - 18% are NIH-funded; many in other countries.
  - 6600 have downloaded more than one version.
  - 2000 citations in scientific journals.
- Petascale biomolecular simulations
  - Collaboration with L.V. Kale (CS).
  - 2002 Gordon Bell award (Phillips *et al.*, SC2002).
  - 2006 target application in NSF Petascale CFP.
  - 2009 PRAC award to prepare for Blue Waters.
- Graphics processor acceleration
  - Early adopters of NVIDIA CUDA technology.
  - Collaborations with Wen-mei Hwu (ECE), IACAT.
  - Stone *et al.*, *J. Comp. Chem.* **28**:2618-2640, 2007



NAMD 2.7b2 (November 12, 2009)

Parallel, object-oriented molecular dynamics code designed for high-performance simulation of large biomolecular systems

Charm++ and Prof. L.V. Kale's Parallel Programming Laboratory

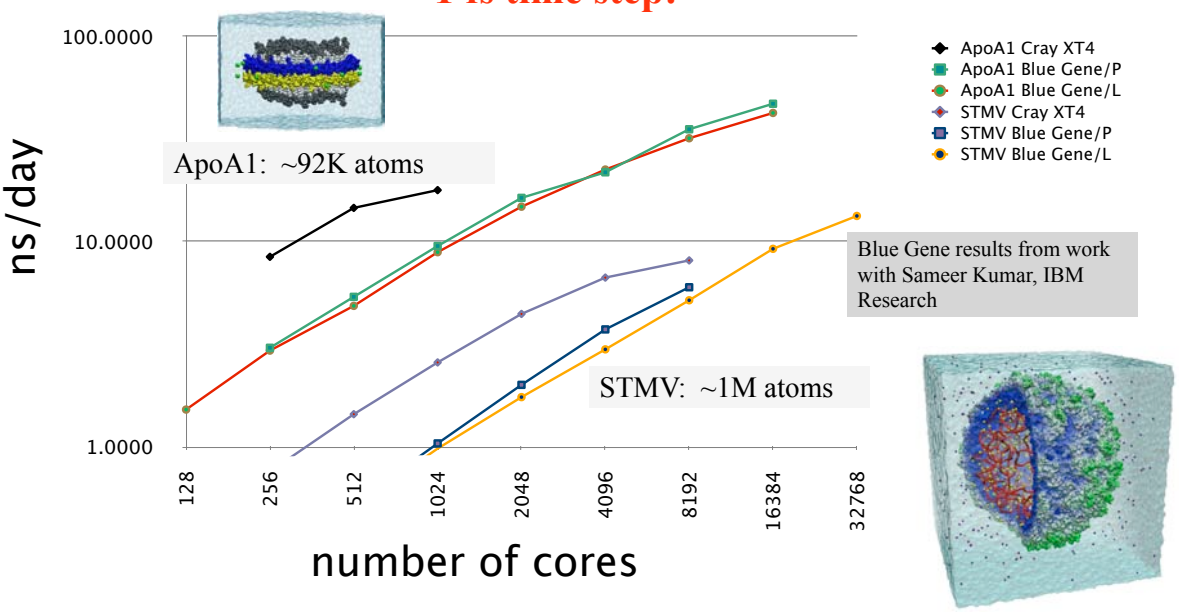
- Distributed free of charge and includes source code.
- Development is supported by the NIH National Center for Research Resources.
- Regular hands-on training; extensive on-line tutorials.
- Sister program VMD.

- New Features:
  1. Collective variable-based calculations
  2. Improved free energy methods for alchemical transformations
  3. Grid-based forces and molecular dynamics flexible fitting
  4. Additional bonded terms for restraining molecular structure
  5. Support for the TIP4 water model
  6. Direct (non-MPI) support for InfiniBand
  7. **NVIDIA CUDA GPU acceleration of non-bonded force evaluation**
  8. **Enhanced performance and scalability (topology awareness)**

NAMD/Charm++ Parallel Scaling Snapshot

Prof. L.V. Kale's Parallel Programming Laboratory

1 fs time step!

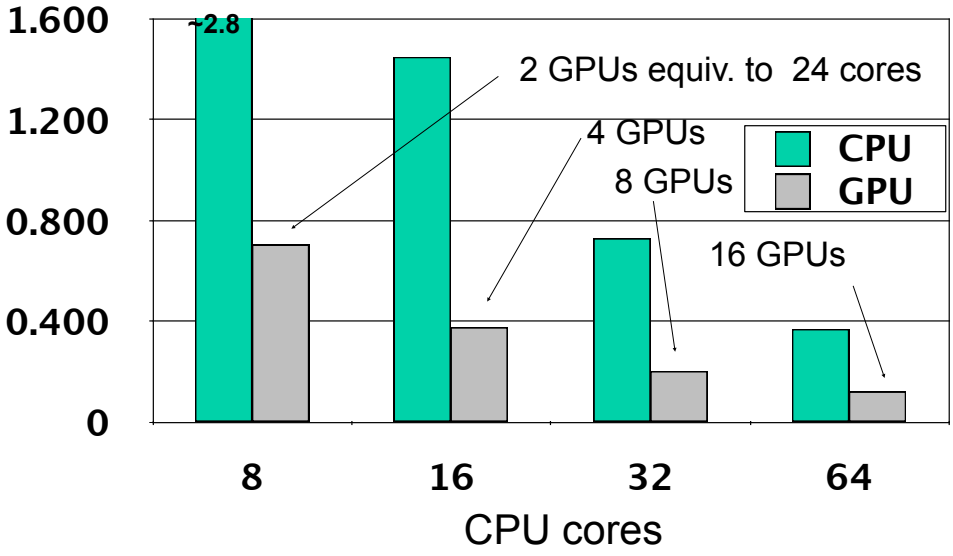


NCSA Lincoln Cluster Performance

(8 Intel cores and 2 NVIDIA Tesla GPUs per node)

STMV (1M atoms) s/step

Phillips, Stone, Schulten, SC2008



Classical Dynamics

at 300K

Energy function: $U(\vec{r}_1, \vec{r}_2, \cdots \vec{r}_N) = U(\vec{R})$

used to determine the force on each atom:

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

yields a set of 3N coupled 2nd-order differential equations that can be propagated forward (or backward) in time.

Initial coordinates obtained from crystal structure, velocities taken at random from Boltzmann distribution.

Maintain appropriate temperature by adjusting velocities.

Classical Dynamics

discretization in time for computing

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

Use positions and accelerations at time t and the positions from time t-δt to calculate new positions at time t+δt.

$$\mathbf{r}(t+\delta t) \approx \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2$$

+

$$\mathbf{r}(t-\delta t) \approx \mathbf{r}(t) - \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2$$

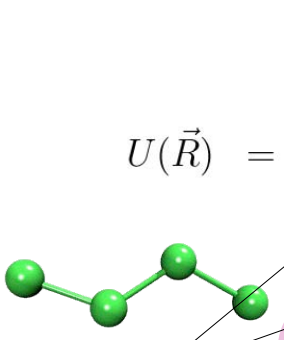
"Verlet algorithm"

$-\vec{\nabla} U(\vec{R}) / m_i$

$$\boxed{\mathbf{r}(t+\delta t) \approx 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \mathbf{a}(t)\delta t^2}$$

Potential Energy Function of Biopolymer

- Simple, fixed algebraic form for every type of interaction.
- Variable parameters depend on types of atoms involved.



heuristic

$$U(\vec{R}) = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} +$$

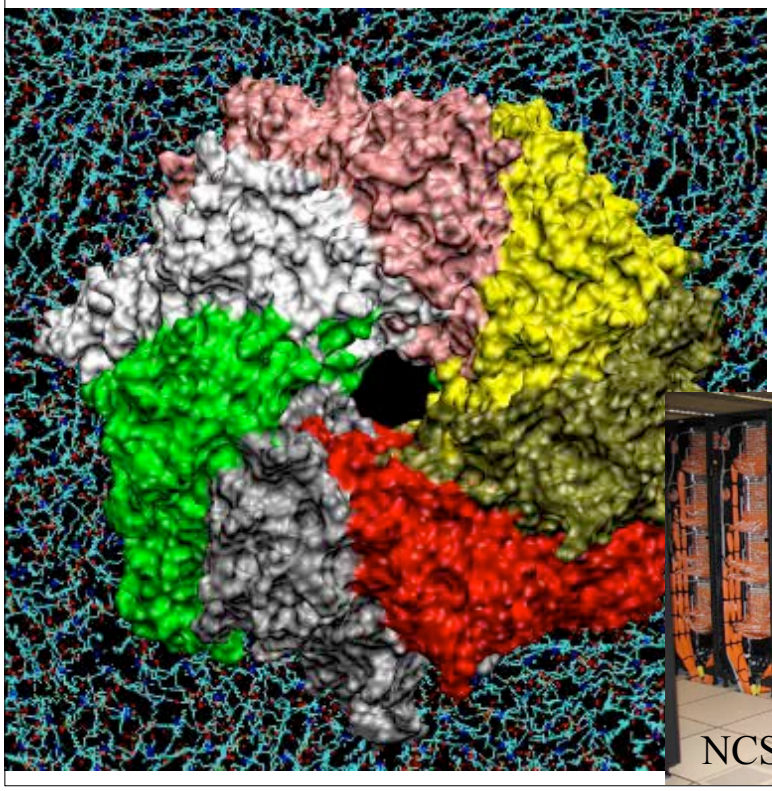
$$\underbrace{\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]}_{U_{dihedral}} +$$

from physics

$$\underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}}$$

Parameters: "force field" like Amber, Charmm; note version number

Large is no problem. But …



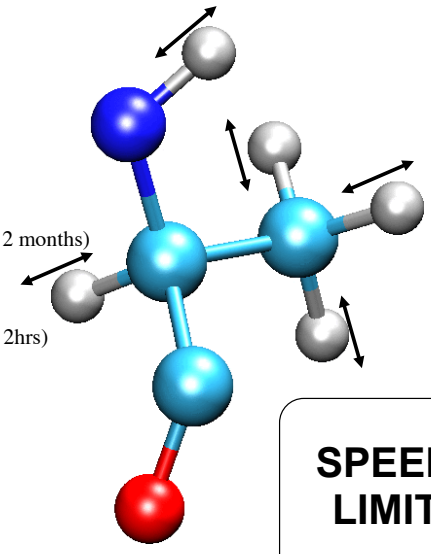
Molecular dynamics simulation of alpha-hemolysin with about 300,000 atoms; 1 million atom simulations are becoming routine today.

NCSA machine room

But long is still a problem!

biomolecular timescale and timestep limits

		steps
Rotation of buried sidechains Local denaturations Allosteric transitions	s	10^{15}
	ms	10^{12} (30 years, 2 months)
	μs	10^{9} (10 days, 2hrs)
Hinge bending Rotation of surface sidechains Elastic vibrations	ns	10^{6} (15 min)
	ps	10^{3}
Bond stretching Molecular dynamics time step →	fs	10^{0}



SPEED LIMIT

δt = 1 fs

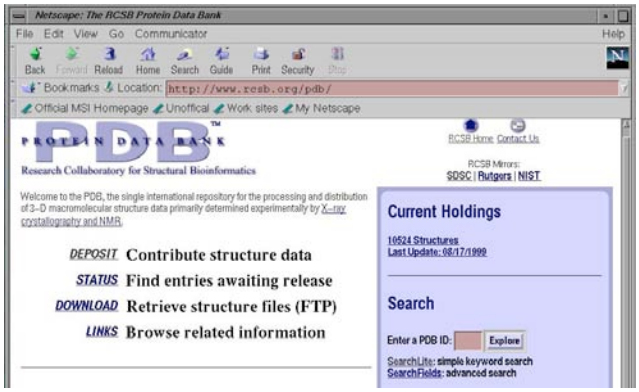
(NSF center, Shaw Res.)

PDB Files

gives one the structure and starting position

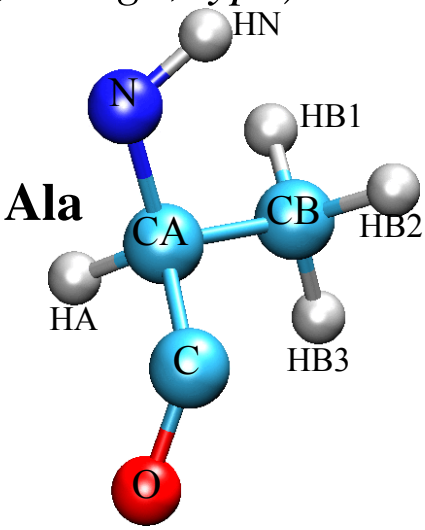
- Simulations start with a crystal structure from the Protein Data Bank, in the standard PDB file format.
- PDB files contain standard records for species, tissue, authorship, citations, sequence, secondary structure, etc.
- We only care about the atom records…
  - atom name (N, C, CA)
  - residue name (ALA, HIS)
  - residue id (integer)
  - coordinates (x, y, z)
  - occupancy (0.0 to 1.0)
  - temp. factor (a.k.a. beta)
  - segment id (6PTI)
- No hydrogen atoms!

  (We must add them ourselves.)



PSF Files

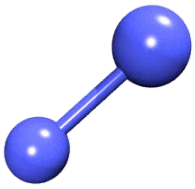
describe atomic properties (mass, charge, type)



- Every atom in the simulation is listed.
- Provides all static atom-specific values:
  - atom name (N, C, CA)
  - atom type (NH1, C, CT1)
  - residue name (ALA, HIS)
  - residue id (integer)
  - segment id (6PTI)
  - atomic mass (in atomic mass units)
  - partial charge (in electronic charge units)
- What is not in the PSF file?
  - coordinates (dynamic data, initially read from PDB file)
  - velocities (dynamic data, initially from Boltzmann distribution)
  - force field parameters (non-specific, used for many molecules)

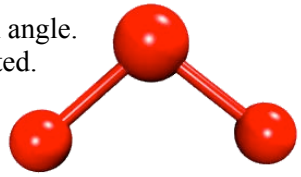
PSF Files

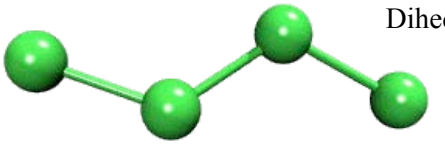
molecular structure (bonds, angles, etc.)



Bonds: Every pair of covalently bonded atoms is listed.

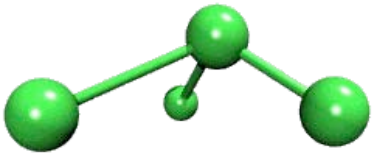
Angles: Two bonds that share a common atom form an angle.
Every such set of three atoms in the molecule is listed.





Dihedrals: Two angles that share a common bond form a dihedral.
Every such set of four atoms in the molecule is listed.

Impropers: Any *planar* group of four atoms forms an improper.
Every such set of four atoms in the molecule is listed.



Preparing Your System for MD

Solvation

Biological activity is the result of interactions between molecules and occurs at the interfaces between molecules (protein-protein, protein-DNA, protein-solvent, DNA-solvent, etc).

mitochondrial bc1 complex

Why model solvation?
- many biological processes occur in aqueous solution
- solvation effects play a crucial role in determining molecular conformation, electronic properties, binding energies, etc

How to model solvation?
- explicit treatment: solvent molecules are added to the molecular system
- implicit treatment: solvent is modeled as a continuum dielectric or so-called implicit force field

Preparing Your System for MD

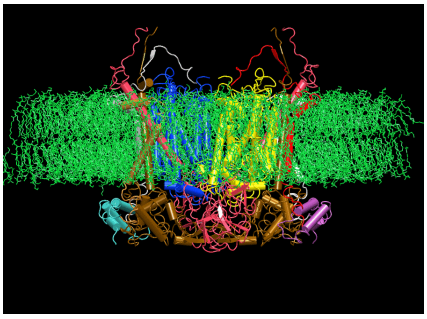
Solvation

Biological activity is the result of interactions between molecules and occurs at the interfaces between molecules (protein-protein, protein-DNA, protein-solvent, DNA-solvent, etc).

Why model solvation?
- many biological processes occur in aqueous solution
- solvation effects play a crucial role in determining molecular conformation, electronic properties, binding energies, etc

How to model solvation?
- explicit treatment: solvent molecules are added to the molecular system
- implicit treatment: solvent is modeled as a continuum dielectric

mitochondrial bc1 complex



Preparing Your System for MD

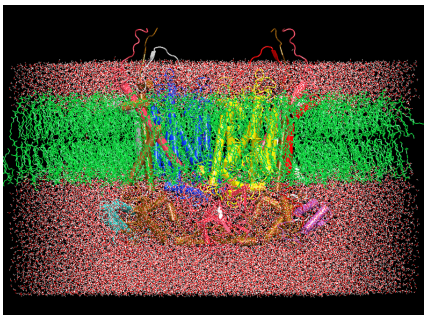
Solvation

Biological activity is the result of interactions between molecules and occurs at the interfaces between molecules (protein-protein, protein-DNA, protein-solvent, DNA-solvent, etc).

Why model solvation?
- many biological processes occur in aqueous solution
- solvation effects play a crucial role in determining molecular conformation, electronic properties, binding energies, etc

How to model solvation?
- explicit treatment: solvent molecules are added to the molecular system
- implicit treatment: solvent is modeled as a continuum dielectric

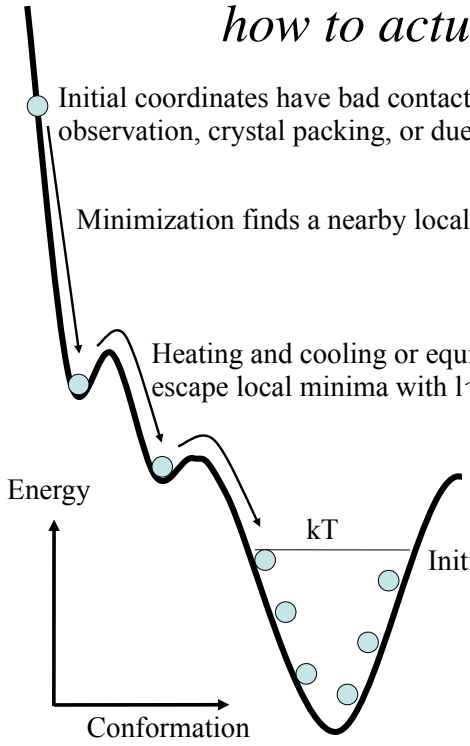
mitochondrial bc1 complex



(Usually periodic! Avoids surface effects)

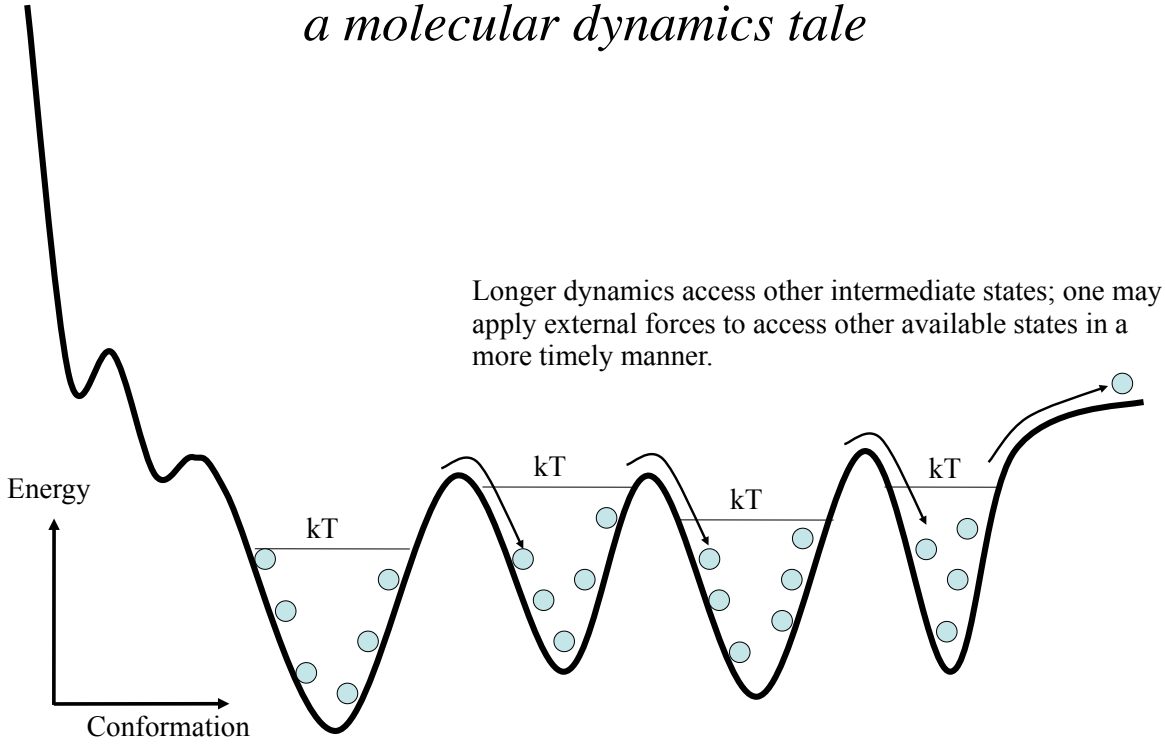
From the Mountains to the Valleys

how to actually describe a protein



From the Mountains to the Valleys

a molecular dynamics tale



Molecular Dynamics Ensembles

Constant energy, constant number of particles (NE)

Constant energy, constant volume (NVE)

Constant temperature, constant volume (NVT)

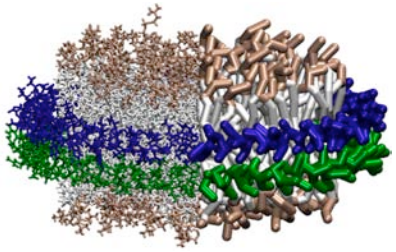
Constant temperature, constant pressure (NPT)

Choose the ensemble that best fits your system and start the simulations, but use NE to check on accuracy of the simulation.

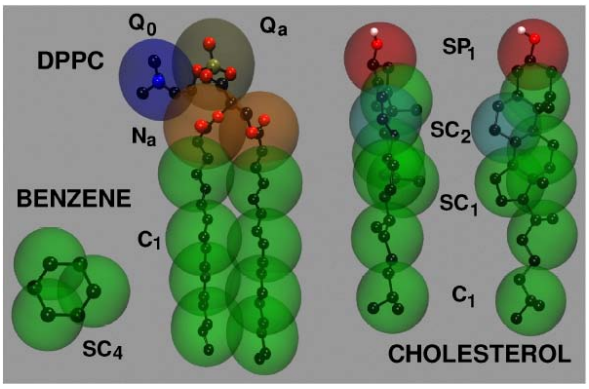
Cutting Corners

cutoffs, PME, rigid bonds, and multiple timesteps

- Nonbonded interactions require order N^2 computer time!
  - Truncating at R_{cutoff} reduces this to order N R_{cutoff}^3
  - Particle mesh Ewald (PME) method adds long range electrostatics at order N log N, only minor cost compared to cutoff calculation.
- Can we extend the timestep, and do this work fewer times?
  - Bonds to hydrogen atoms, which require a 1fs timestep, can be held at their equilibrium lengths, allowing 2fs steps.
  - Long range electrostatics forces vary slowly, and may be evaluated less often, such as on every second or third step.
- Coarse Graining

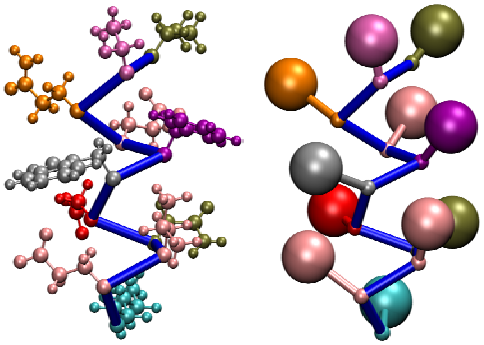


Residue-Based Coarse-Grained Model



- Lipid model: MARTINI
- Level of coarse-graining: ~4 heavy atoms per CG bead
- Interactions parameterized based on experimental data and thermodynamic properties of small molecules

- Protein model uses two CG beads per residue
- One CG bead per side chain another for backbone

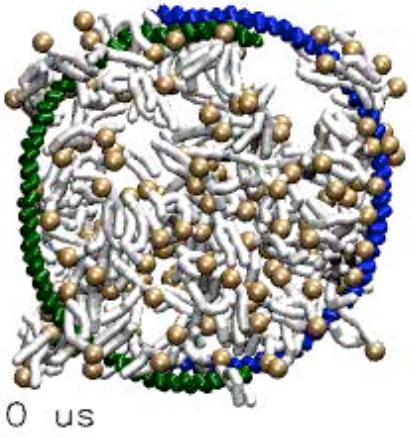


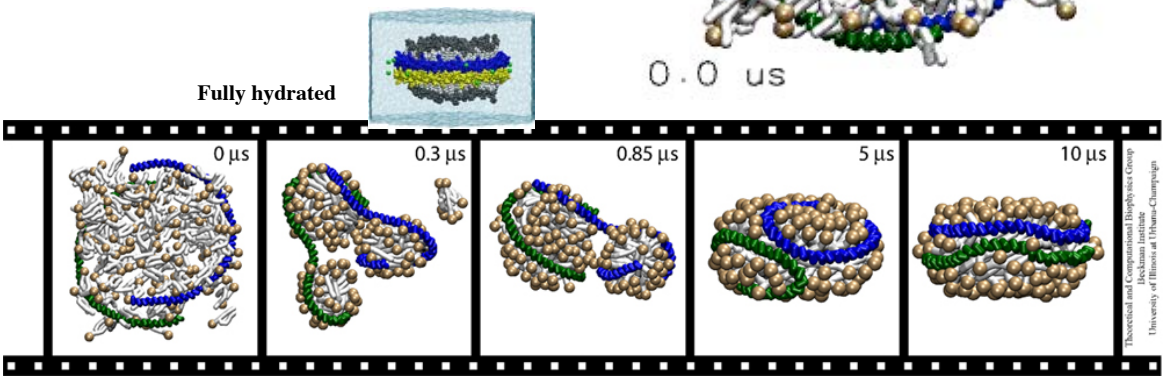
All-atom peptide CG peptide

Peter L. Freddolino, Anton Arkhipov, Amy Y. Shih, Ying Yin, Zhongzhou Chen, and Klaus Schulten. **Application of residue-based and shape-based coarse graining to biomolecular simulations.** In Gregory A. Voth, editor, *Coarse-Graining of Condensed Phase and Biomolecular Systems*, chapter 20, pp. 299-315. Chapman and Hall/CRC Press, Taylor and Francis Group, 2008.

Nanodisc Assembly CG MD Simulation

- 10 μs simulation
- Assembly proceeds in two steps:
  - Aggregation of proteins and lipids driven by the hydrophobic effect
  - Optimization of the protein structure driven by increasingly specific protein-protein interactions
- Formation of the generally accepted double-belt model for discoidal HDL

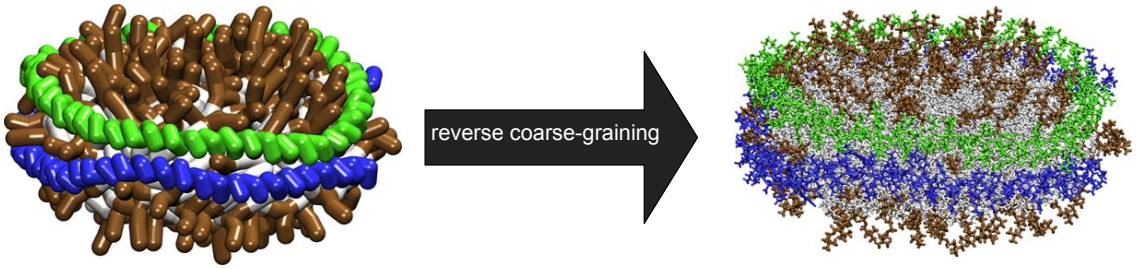




A. Shih, A. Arkhipov, P. Freddolino, and K. Schulten. *J. Phys. Chem. B*, 110:3674–3684, 2006; A. Shih, P. Freddolino, A. Arkhipov, and K. Schulten. *J. Struct. Biol*., 157:579–592,2007; A. Shih, A. Arkhipov, P. Freddolino, S. Sligar, and K. Schulten. *Journal of Physical Chemistry B*, 111: 11095 - 11104, 2007; A. Shih, P. Freddolino, S. Sligar, and K. Schulten. *Nano Letters*, 7:1692-1696, 2007.

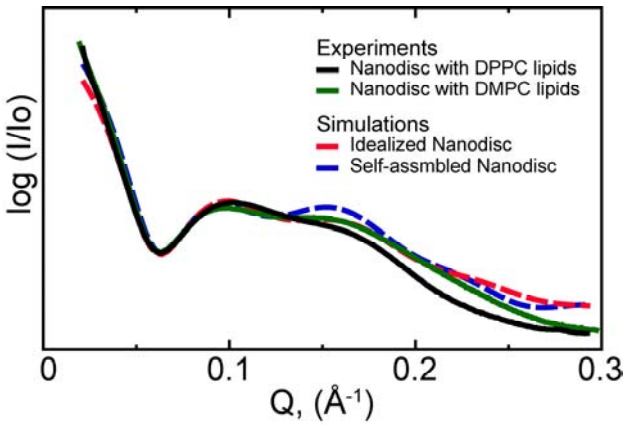
Validation of Simulations

reverse coarse-graining and small-angle X-ray scattering

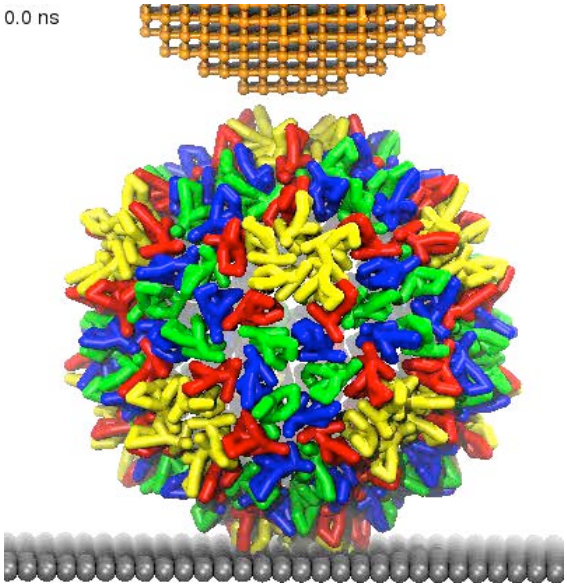


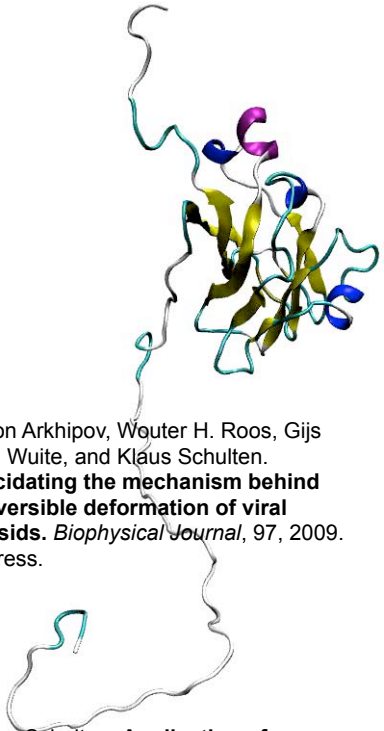
Reverse coarse-graining:
1. Map center of mass of the group of atoms represented by a single CG bead to that beads location
2. MD minimization, simulated annealing with restraints, and equilibration to get all-atom structure

Small-angle X-ray scattering:
Calculated from reverse coarse-grained all-atom model and compared with experimental measurements



Shape-Based Coarse-Grained (CG) model



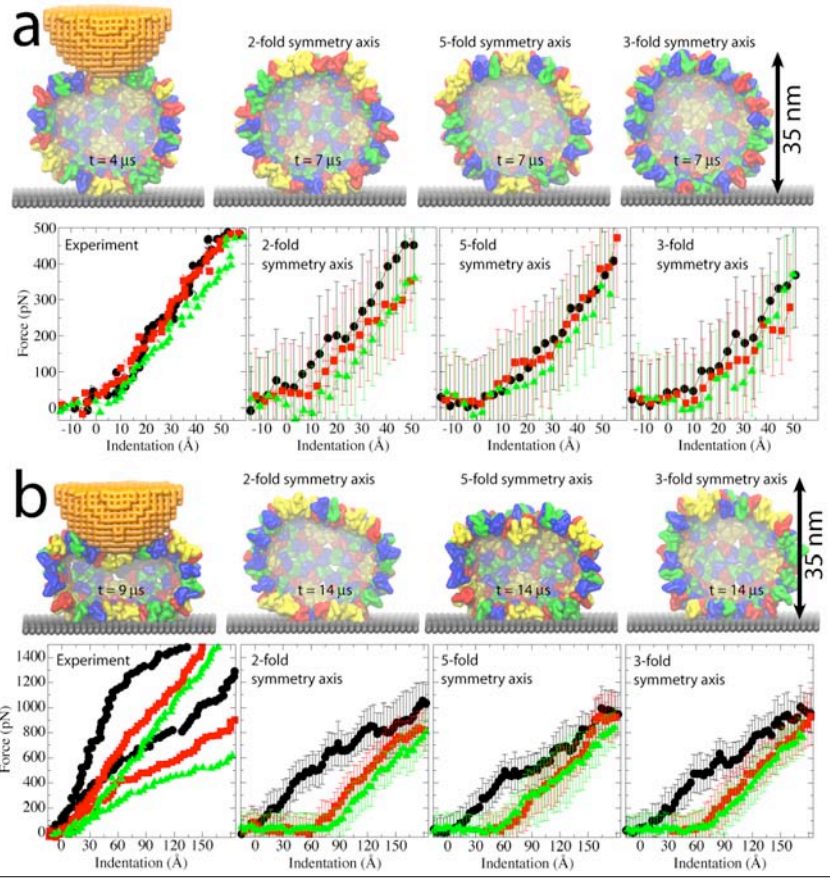


Anton Arkhipov, Wouter H. Roos, Gijs J. L. Wuite, and Klaus Schulten. **Elucidating the mechanism behind irreversible deformation of viral capsids.** *Biophysical Journal*, 97, 2009. In press.

- **Fully automatic**
- **Number of CG beads is chosen by a user (we used ~200 atoms per CG bead)**

Peter L. Freddolino, Anton Arkhipov, Amy Y. Shih, Ying Yin, Zhongzhou Chen, and Klaus Schulten. **Application of residue-based and shape-based coarse graining to biomolecular simulations.** In Gregory A. Voth, editor, *Coarse-Graining of Condensed Phase and Biomolecular Systems*, chapter 20, pp. 299-315. Chapman and Hall/CRC Press, Taylor and Francis Group, 2008.

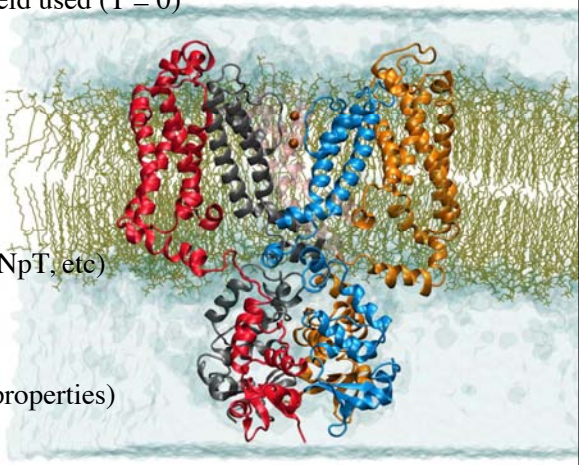
Reversible and irreversible indentations



Anton Arkhipov, Wouter H. Roos, Gijs J. L. Wuite, and Klaus Schulten. **Elucidating the mechanism behind irreversible deformation of viral capsids.** *Biophysical Journal* 97: 2061-2069 (2009).

Summary: Steps in a Typical MD Simulation

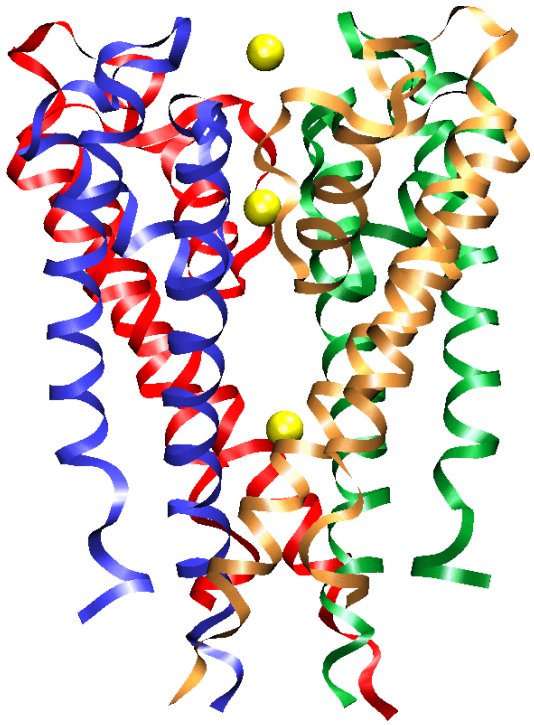
- 1. Prepare molecule
  - Read in pdb and psf file
  - Usually requires setting up the system, e.g., solvation
  - Many tools available in VMD
- 2. Minimization
  - Reconcile observed structure with force field used (T = 0)
- 3. Heating
  - Raise temperature of the system
- 4. Equilibration
  - Ensure system is stable
- 5. Dynamics
  - Simulate under desired conditions (NVE, NpT, etc)
  - Collect your data
- 6. Analysis
  - Evaluate observables (macroscopic level properties)
  - Or relate to single molecule experiments
  - Many tools available in VMD



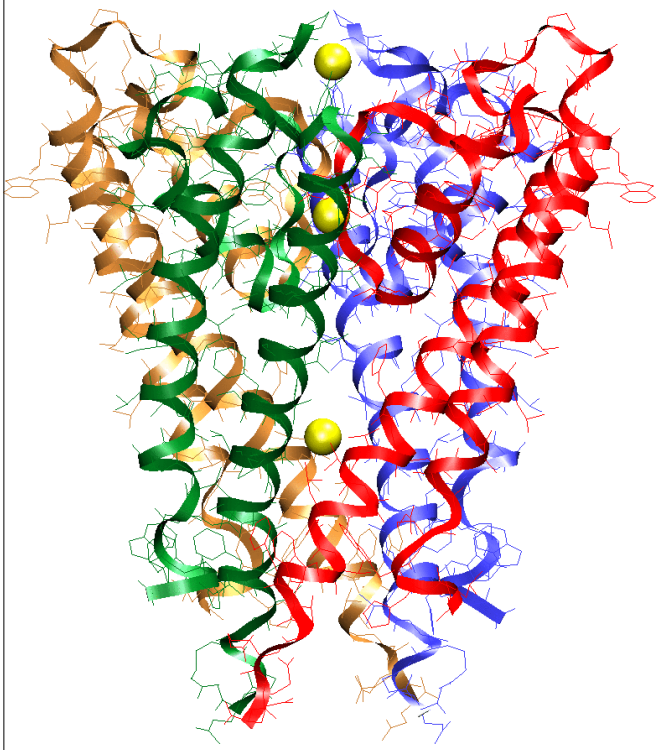
Example: MD Simulations of the K^+ Channel Protein

Ion channels are membrane - spanning proteins that form a pathway for the flux of inorganic ions across cell membranes.

Potassium channels are a particularly interesting class of ion channels, managing to distinguish with impressive fidelity between K^+ and Na^+ ions while maintaining a very high throughput of K^+ ions when gated.

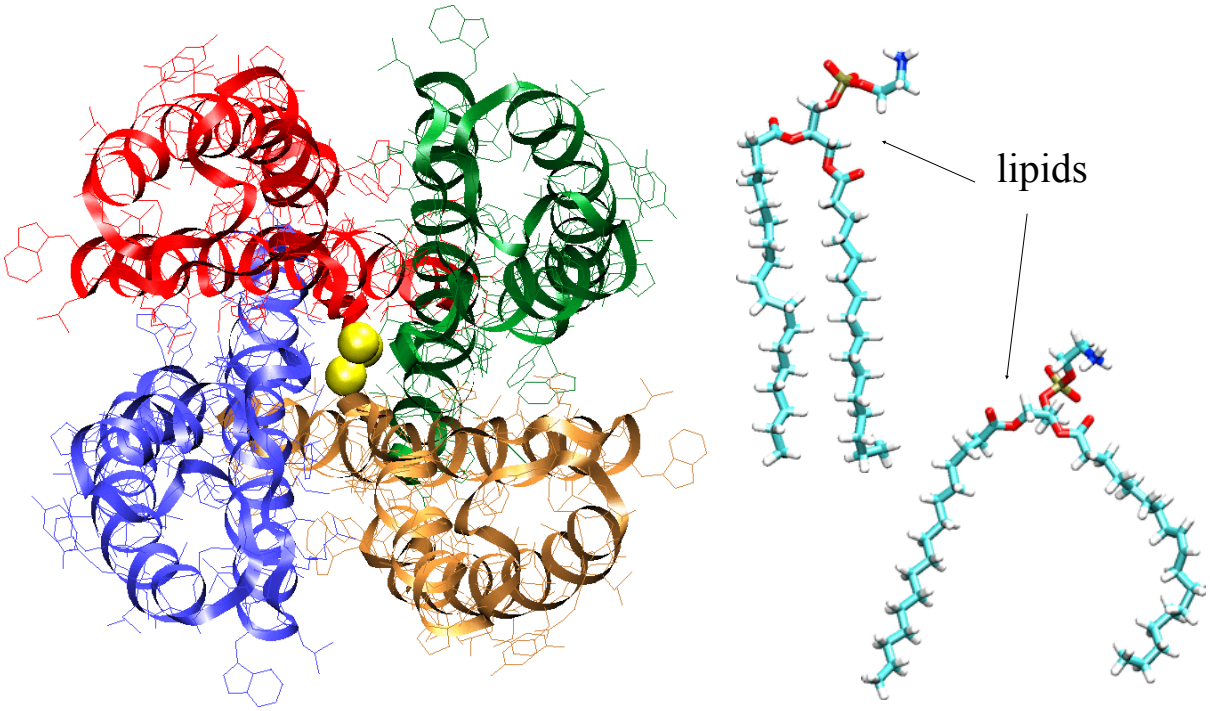


Setting up the system (1)



- retrieve the PDB (coordinates) file from the Protein Data Bank
- add hydrogen atoms using PSFGEN
- use psf and parameter files to set up the structure; needs better description than available in Charmm to account for ion selectivity
- minimize the protein structure using NAMD2

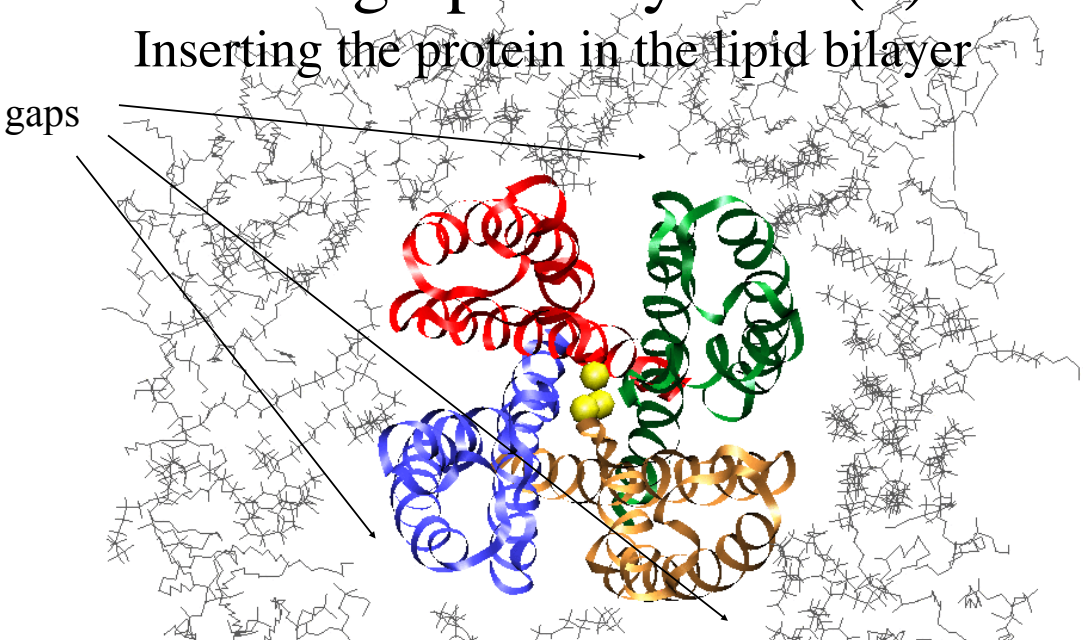
Setting up the system (2)



Simulate the protein in its natural environment: solvated lipid bilayer

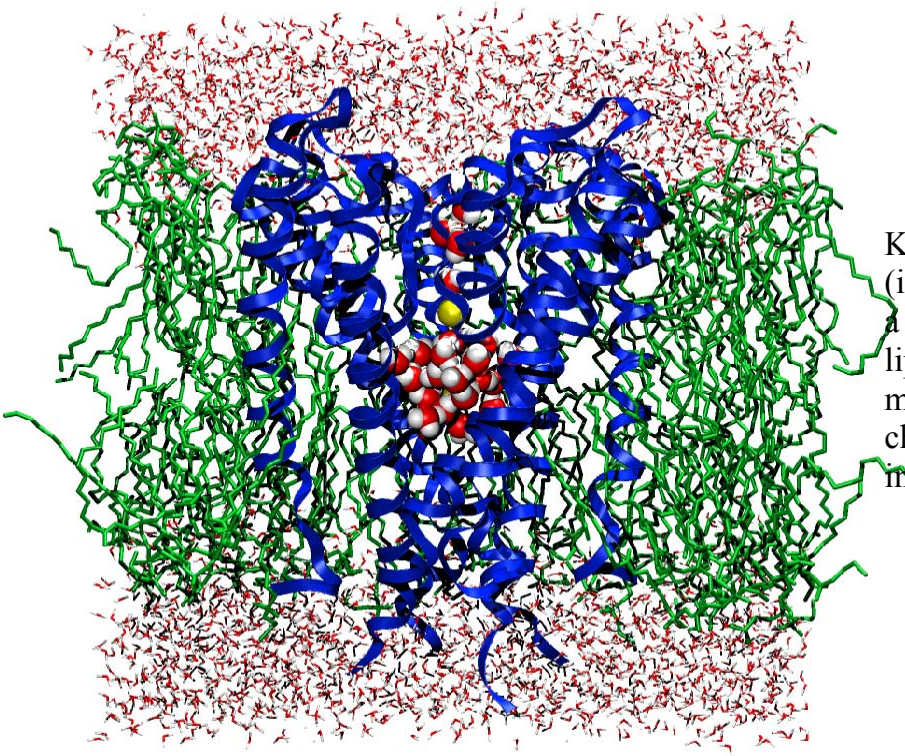
Setting up the system (3)

Inserting the protein in the lipid bilayer



Automatic insertion into the lipid bilayer leads to big gaps between the protein and the membrane => long equilibration time required to fill the gaps. Solution: manually adjust the position of lipids around the protein. Employ constant (lateral and normal) pressure control.

The system



solvent

Kcsa channel protein (in blue) embedded in a (3:1) POPE/POPG lipid bilayer. Water molecules inside the channel are shown in vdW representation.

solvent

Simulating the system: Free MD

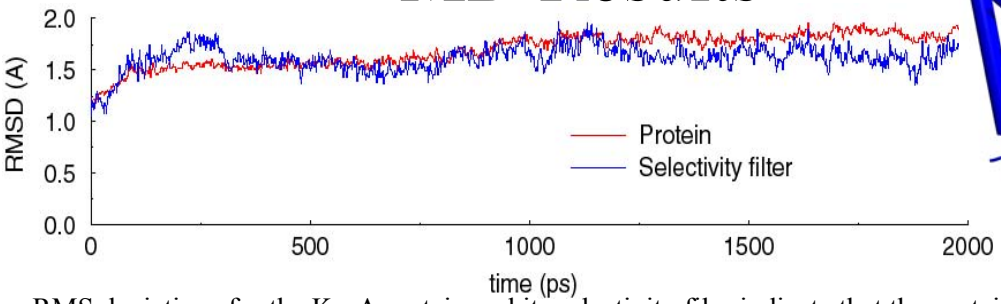
Summary of simulations:

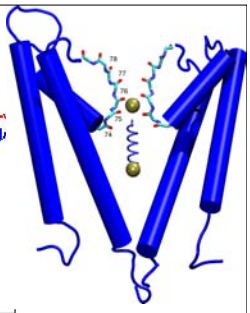
- protein/membrane system contains 38,112 atoms, including 5117 water molecules, 100 POPE and 34 POPG lipids, plus K^+ counterions
- CHARMM26 forcefield
- periodic boundary conditions, PME electrostatics
- 1 ns equilibration at 310K, NpT
- 2 ns dynamics, NpT

Program: NAMD2

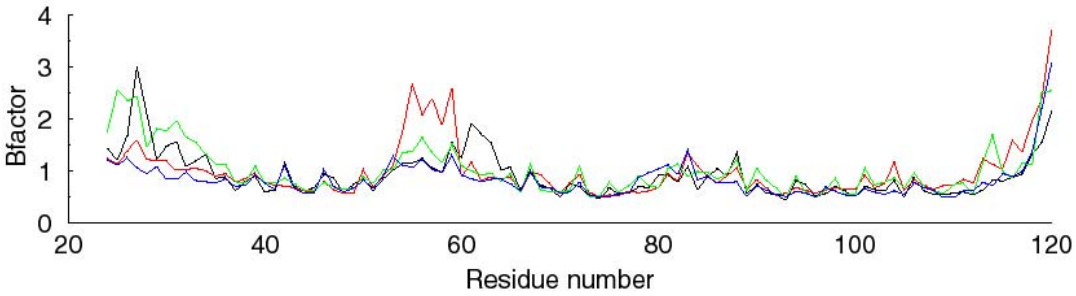
Platform: Cray T3E (Pittsburgh Supercomputer Center) or local computer cluster; choose ~1000 atoms per processor.

MD Results



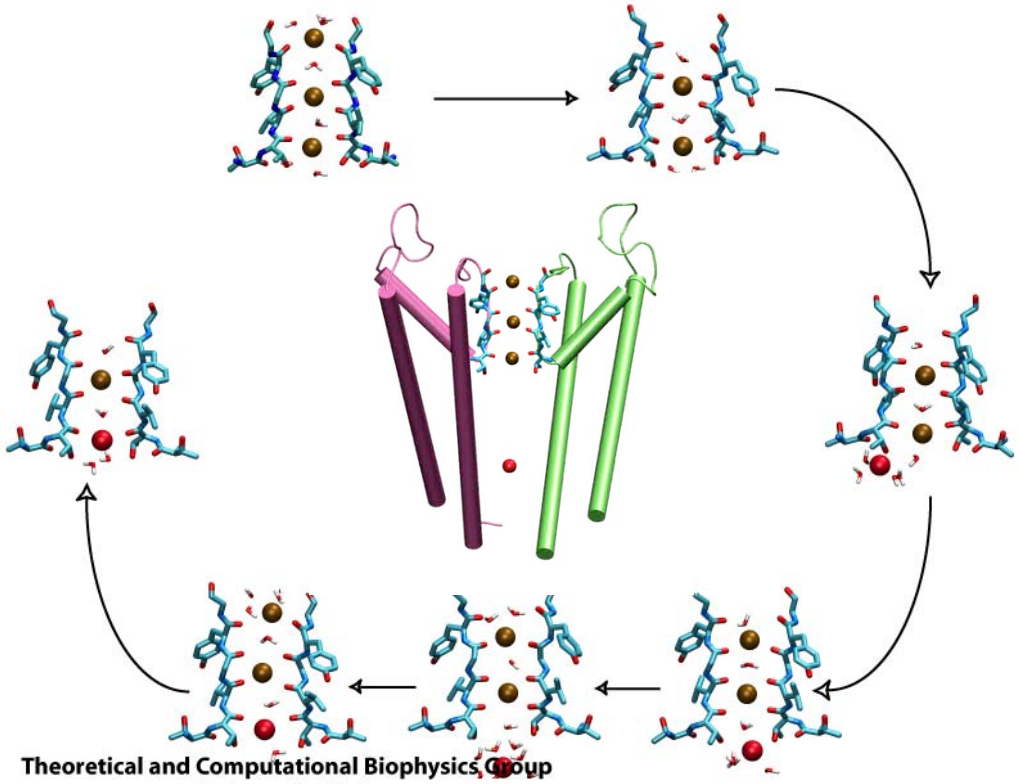


RMS deviations for the KcsA protein and its selectivity filer indicate that the protein is stable during the simulation with the selectivity filter the most stable part of the system.



Temperature factors for individual residues in the four monomers of the KcsA channel protein indicate that the most flexible parts of the protein are the N and C terminal ends, residues 52-60 and residues 84-90. Residues 74-80 in the selectivity filter have low temperature factors and are very stable during the simulation.

Simulation of Ion Conduction (here for Kv1.2)



Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

Theoretical and Computational Biophysics Group Developers

NIH Center for Research Resources

Funding: NIH, NSF

- focus on systems biology
- focus on quantum biology
- theoretical biophysics
- computational biophysics
- develops renewable energy
- guides bionanotechnology