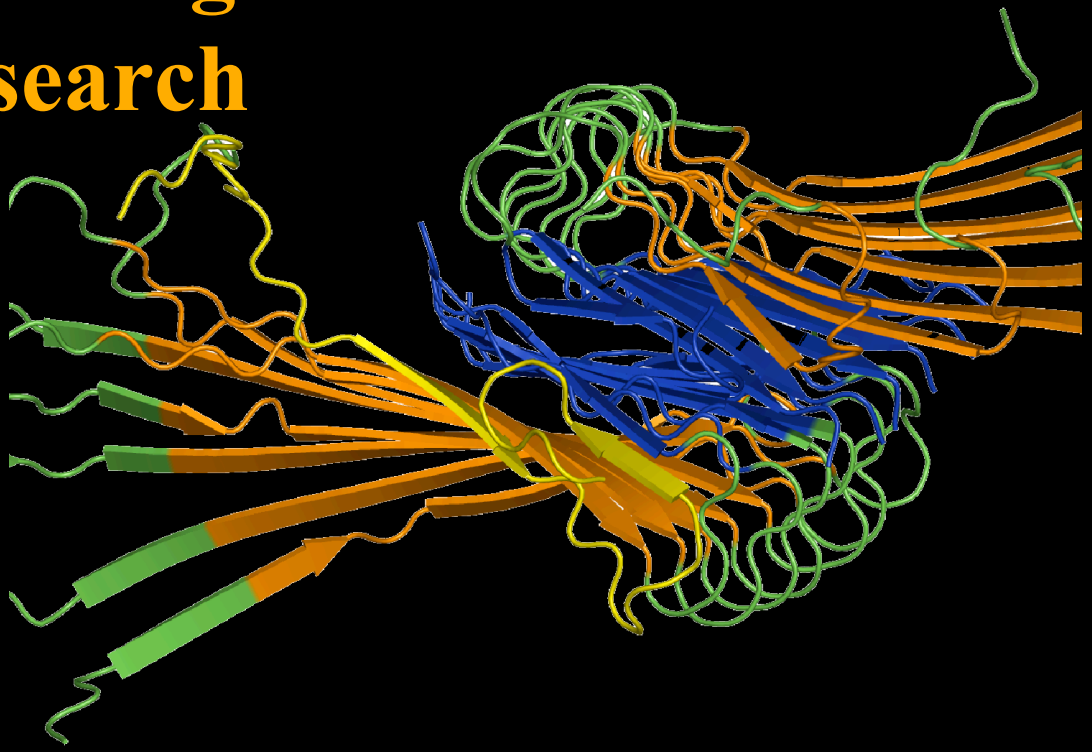
Large Scale Production Computing Requirements for Biological and Environmental Research



Teresa Head-Gordon
Lawrence Berkeley National Laboratory
UC Berkeley

NERSC_BER'09

Molecular Theory and Simulation

Every attempt to employ mathematical methods in the study of (bio)chemical questions must be considered profoundly irrational and contrary to the spirit of (bio)chemistry. If mathematical analysis should ever hold a prominent place in (bio)chemistry - an aberration which is happily almost impossible - it would occasion a rapid and widespread degeneration of that science.

A Comte (1830)

Molecular Theory and Simulation

Every attempt to employ mathematical methods in the study of (bio)chemical questions must be considered profoundly irrational and contrary to the spirit of (bio)chemistry. If mathematical analysis should ever hold a prominent place in (bio)chemistry - an aberration which is happily almost impossible - it would occasion a rapid and widespread degeneration of that science.

A Comte (1830)

The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of (bio)chemistry are thus completely known, and the difficulty lies only in the fact that the exact application of these laws leads to equations much too complicated to be soluble.

P. A. M. Dirac (1929)

Theoretical Framework for Molecular Simulation

In the chemical energy regime, where most of molecular materials, chemistry, biology operates, we organize our theoretical framework into four basic theories:

(1) Quantum Mechanics ⟷ **Potential energy surfaces**

QM allows prediction of potential energy surfaces on which atoms move via the Born-Oppenheimer approximation

Theoretical Framework for Molecular Simulation

In the chemical energy regime, where most of molecular materials, chemistry, biology operates, we organize our theoretical framework into four basic theories:

(1) Quantum Mechanics ⟷ Potential energy surfaces

QM allows prediction of potential energy surfaces on which atoms move via the Born-Oppenheimer approximation

(2) Classical Mechanics ⟷ Moving on PE surfaces

Atoms and molecules obey classical motion on these potential energy surfaces; i.e. Newtons' equations of motion: F=ma (QM or MModel)

(1) and (2) describes physical matter at level of microscopic atoms and molecules

Theoretical Framework for Molecular Simulation

(3) Thermodynamics ⟷ Macroscopic Observables

Measures equilibrium properties at the level of macroscopic observables under externally controllable conditions: temperature, pressure, etc

(4) Statistical Mechanics ⟷ Microscopic to macroscopic

Statistical Mechanics permits correct *statistical* averaging of molecular info based on distributions that depend on externally controllable specified conditions (T, P, etc), to connect microscopic theories to macroscopic observables.

It recognizes "importance sampling" through stochastic Monte Carlo simulations

Statistical Mechanics/Numerical Simulation

When statistical mechanics theories can not be expressed in analytical form, then numerical methods are necessary to simulate the laws of statistical mechanics. Good sampling statistics are necessary to converge observables

The statistical distributions we sample all depend on the potential energy (QM). Potential energy surfaces are the basic physical interactions between molecules, atoms, or even the basic constituents of atoms, such as electrons. We can model them at different levels of accuracy

Thus we must analyze the trade off between accuracy and tractability of potential energy surfaces vs. sampling statistics (MD or MC)

Models, Algorithms, Hardware

Some community efforts, but also alot of "roll-your-own"!

Models: depends on science question (better developed)

- most use fixed charge models O(NlogN)✓
- polarizable FFs and ab initio MD [cO(NlogN) and $CN^{2.5\text{-}3.0}$]
- get rid of water! (Folding@home)
- get rid of atomic detail of atoms!

Methods/Algorithms: Enhance sampling

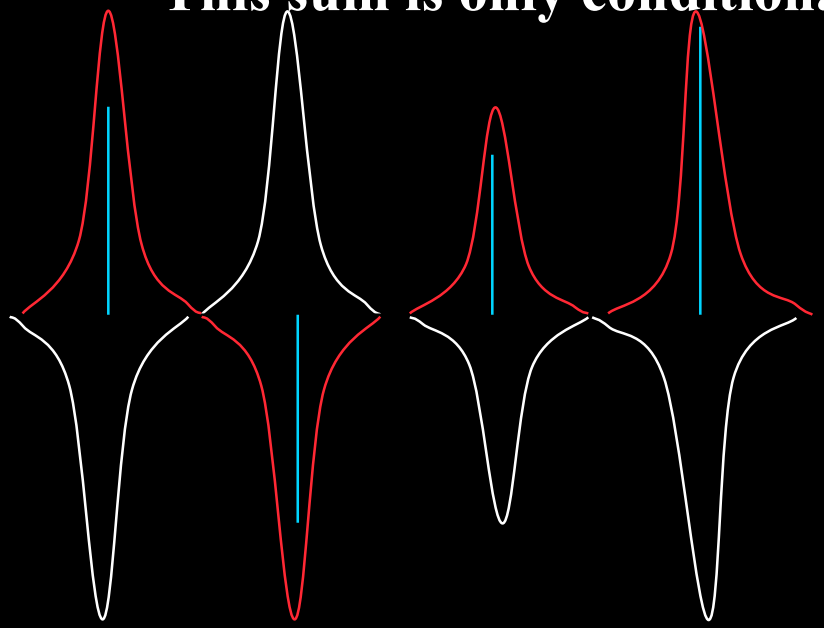
- increase effective time step of atomistic MD (under-developed)
- Enhanced MC techniques such as Replica Exchange ✓ (active)

Hardware Implementations: Sampling and model

- parallelization ✓(under-developed for new advanced models)
- get a better computer! ✓ (BGene, DE Shaw, Fold@home, GPUs)

Long-Ranged Forces

This sum is only conditionally convergent (depending on the order in which you add the terms).



$$V_{tot} = \sum_{i}\sum_{j}\sum_{\vec{n}}{}' \frac{q_i q_j}{\left|\vec{r}_{ij} + L\vec{n}\right|}$$

$$\vec{n} = \left(n_x L, n_y L, n_z L\right)$$

Reformulate original non-convergent sum with two sums: a real-space sum (r-sum: screened) and inverse-space sum (k-sum: compensating) which we can derive from Poisson's equation

$$V_{qq} = \sum_{i>j}^{N}\left(\sum_{|\mathbf{n}|=0}^{\infty} q_i q_j \frac{erfc\left(\kappa\left|\mathbf{r}_{ij}+\mathbf{n}\right|\right)}{\left|\mathbf{r}_{ij}+\mathbf{n}\right|} + \frac{1}{\pi L^3}\sum_{\mathbf{k}\neq\mathbf{0}} q_i q_j \frac{4\pi^2}{k^2}\exp\left(-k^2/4\kappa^2\right)\cos\left(\mathbf{k}\cdot\mathbf{r}_{ij}\right)\right) + V_{self}$$

Long-Ranged Forces

The inverse space part is evaluated with a Fast Fourier transform evaluated on a discrete lattice in space

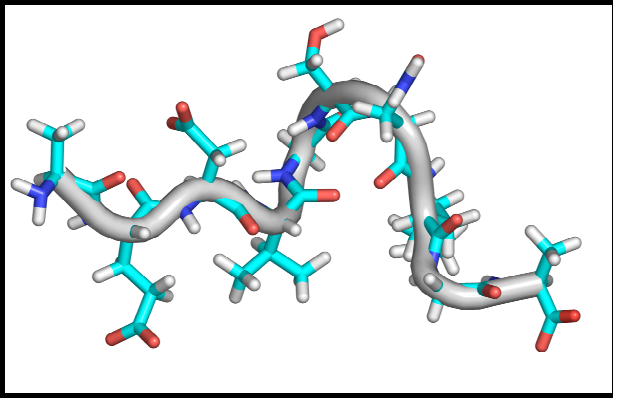
$$V_{qq} = \sum_{i>j}^{N} \left(\sum_{|\mathbf{n}|=0}^{\infty} q_i q_j \frac{erfc\left(\kappa \left|\mathbf{r}_{ij} + \mathbf{n}\right|\right)}{\left|\mathbf{r}_{ij} + \mathbf{n}\right|} + \frac{1}{\pi L^3} \sum_{\mathbf{k}\neq\mathbf{0}} q_i q_j \frac{4\pi^2}{k^2} \exp\left(-k^2/4\kappa^2\right) \cos\left(\mathbf{k}\cdot\mathbf{r}_{ij}\right) \right) + V_{self}$$

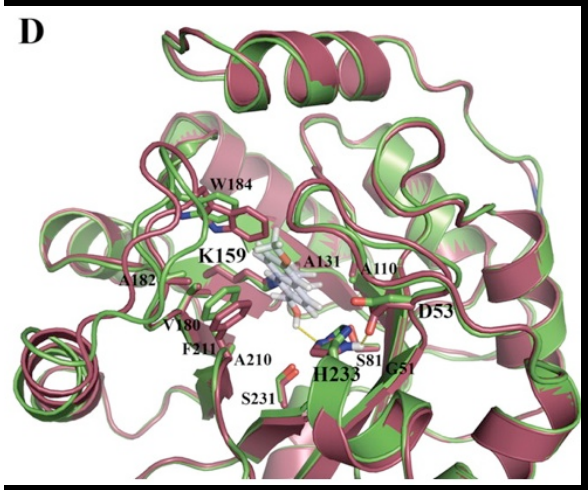
- **Conventional algorithm scales as $N^{3/2}$**
- **Particle Mesh Ewald O(NlogN)**
  Spatial Decomposition in r-space; Parallelization of FFT's in k-space
- **Evaluate Ewald in r-space using FMM techniques O(N). Hard to beat PME for continuous potentials and smooth densities**

This rate limiting step must be evaluated many times due to symplectic structure of MD ($10^{6\text{-}9}$). In the next 5 years it would be 10^9 to 10^{15} times per system (of many, many systems)

Big Science Opportunities in Simulation

Genome efforts: ~1/3 of the sequences in the human genome involve unstructured proteins!



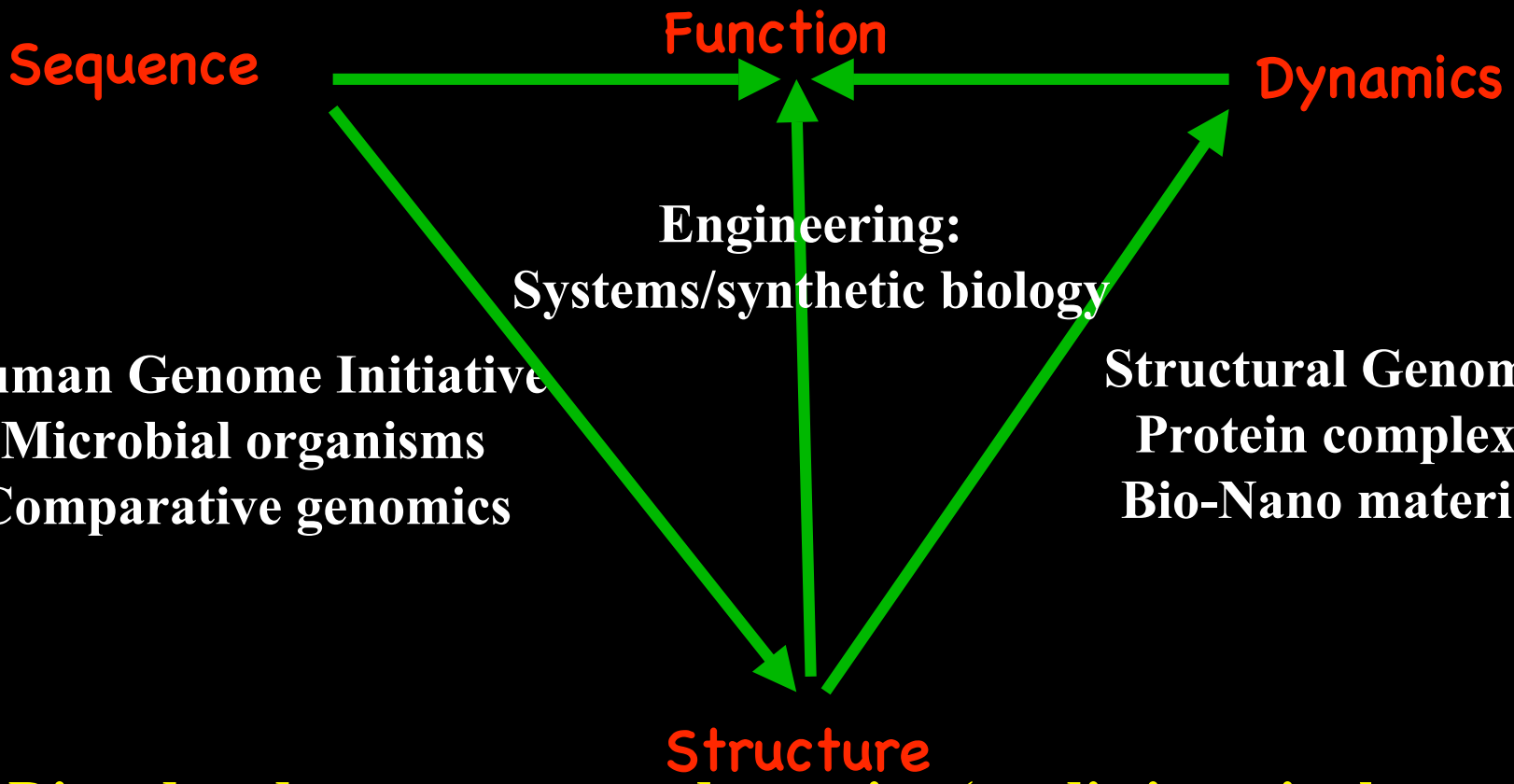


Biofuels: MD simulation studies of existing cellulases or de novo design of new enzymes

Disordered Proteins

Functional Annotation Initiative
Yeast two-hybrid screening
Gene expression micro-arrays

Single Molecule
Temporal measurements
Force-velocity



Human Genome Initiative
Microbial organisms
Comparative genomics

Structural Genomics
Protein complexes
Bio-Nano materials

Disordered or unstructured proteins (no distinct single tertiary structure) comprise 1/3 of the sequences in the human genome!

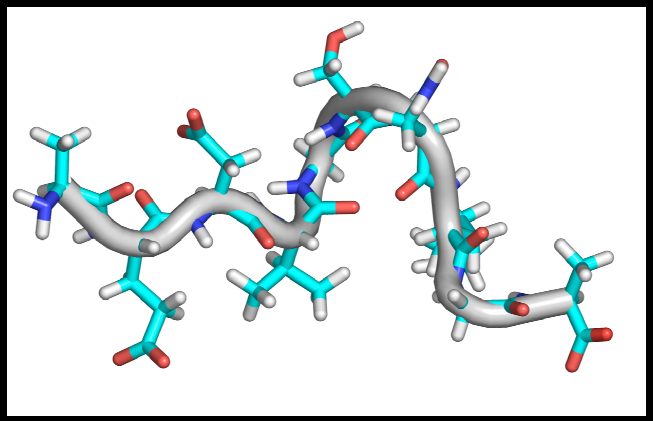
Disordered Proteins

Traditional experimental x-ray determination of (functionally important) unstructured peptides and proteins is a non-starter.

NMR is the obvious experiment but is a challenge to interpret due to diversity of populations. NMR can provide restraints on the population profile but can't quantify populations

By contrast, simulations generate fully detailed ensembles and populations, but accuracy may be questionable due to the empirical nature of the potential energy description and issues of convergence to equilibrium.

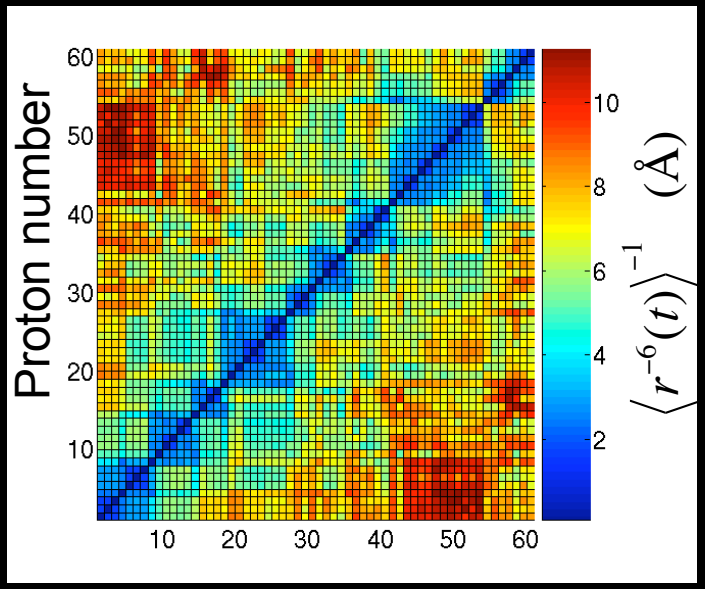
Bioinformatics and Genomics can't provide the solution to this problem

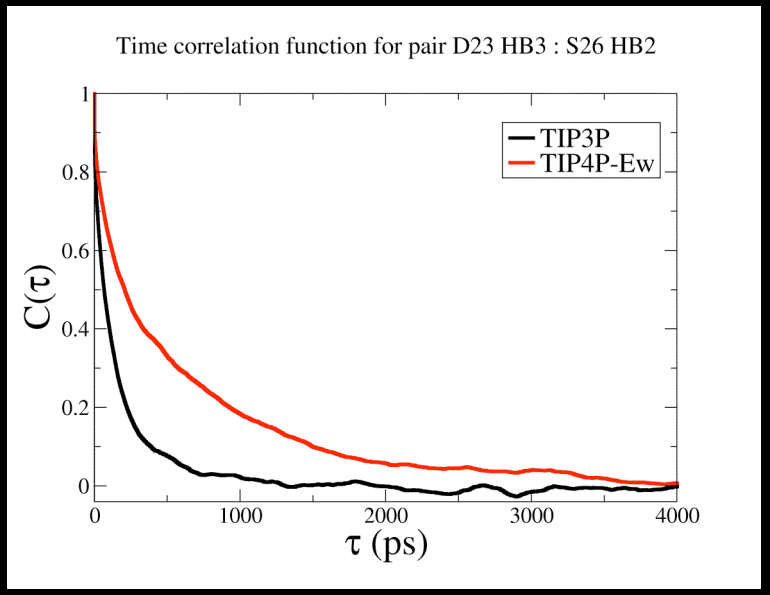


Disordered Proteins

The physics of NMR is very well understood:

Simulation can simulate the experimental observable!





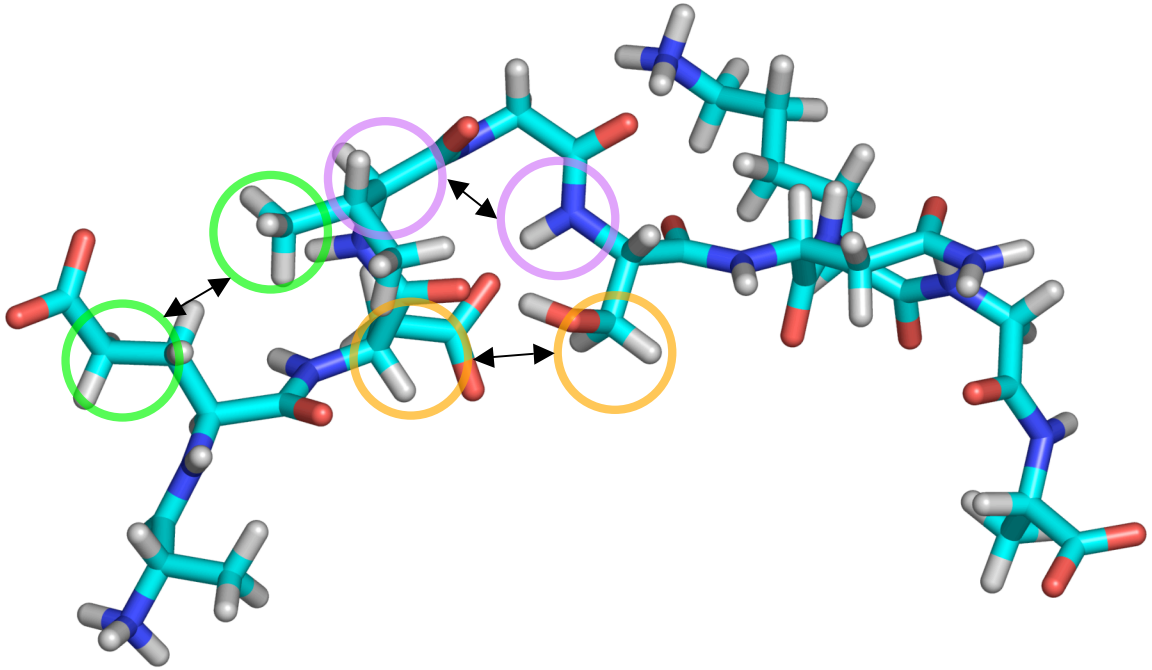
Simulated ensemble distances + Simulated ensemble dynamics

$$C(\tau) = \left\langle r^{-6}(t)\right\rangle^{-1} \left\langle \frac{P_2(\cos\chi_{t,t+\tau})}{r^3(t)r^3(t+\tau)} \right\rangle$$

$P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}$ and $\chi_{t,t+\tau}$ = angle between the interspin vector at t and $t+\tau$

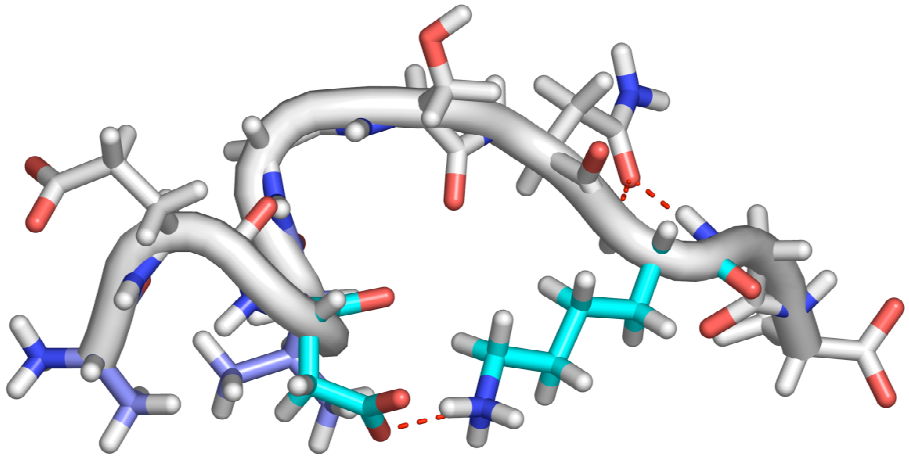
$$J(\omega) = \int_{-\infty}^{\infty} C(\tau)\, e^{-i\omega\tau} d\tau$$

Disordered Proteins



Simulation necessary to define structure of disordered proteins!

Need to think about the possibility of computational beamlines of annotating this 1/3 of genome



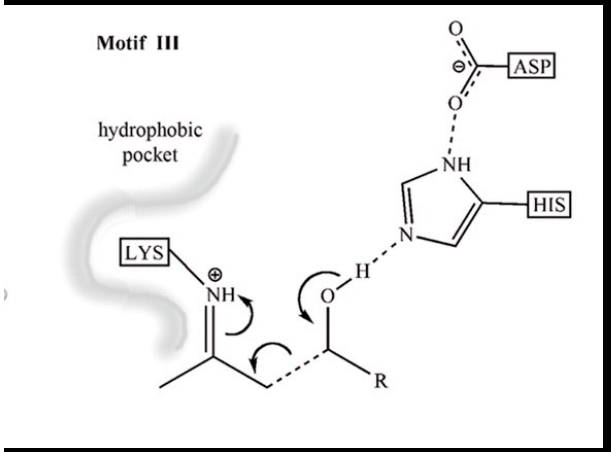
Biofuels and De Novo Enzyme Design

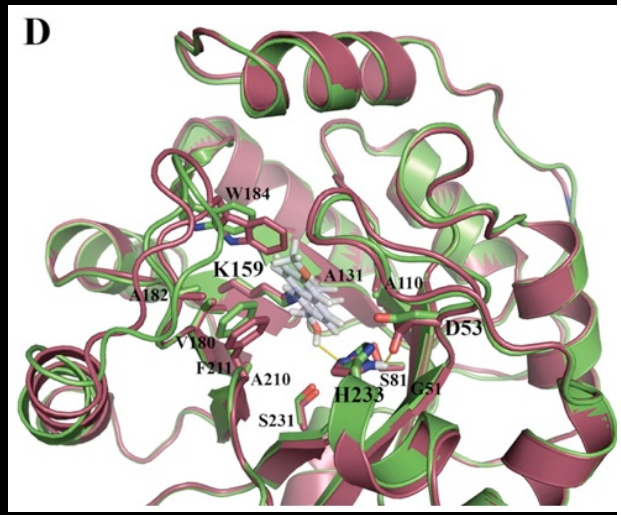
❖ Computational de novo design approach emphasizes catalysis of virtually any reaction

- ❖Non-natural substrates
- ❖Substrate, catalytic, product promiscuity and regulation

❖We wish to extend that goal to virtually any reaction under virtually any solution condition for arbitrary scaffold:

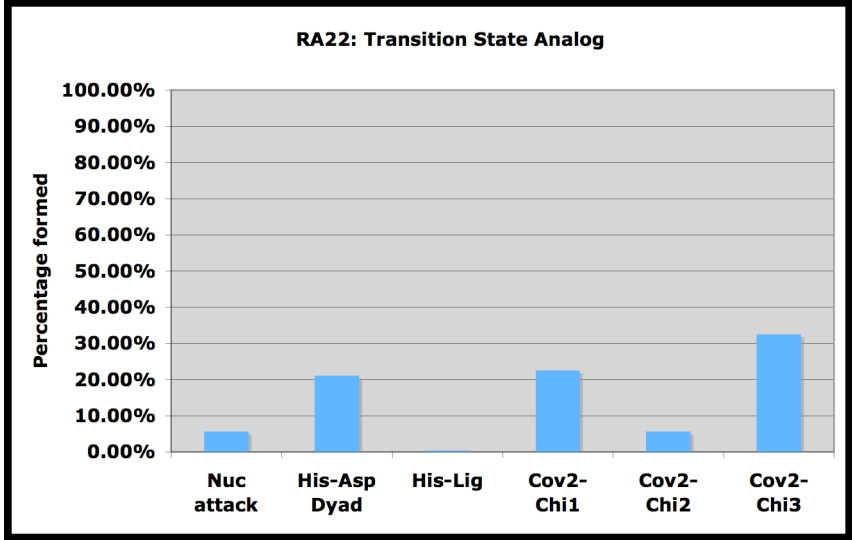
- ❖Temperature, salt, pH
- ❖Non-natural solvents such as ionic liquids
- ❖Native cellular environments

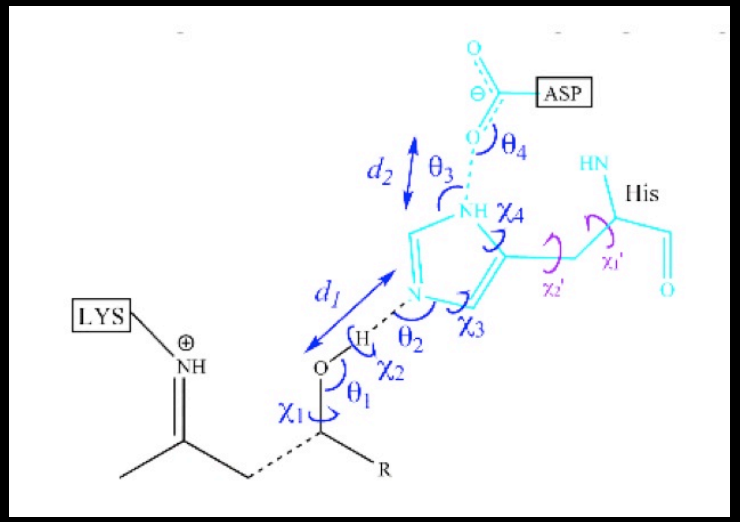




Biofuels and De Novo Enzyme Design

While statics met, temporal criteria fails completely. This is most likely source of poor catalysis in de novo design.





Must combine rational design with *computational* directed evolution: we want high-throughput and therefore fast computational *dynamical* assays of activity: increases complexity multi-fold over static approaches

Challenges and Opportunities in Molecular Bio/Chemical Simulation

Primary issues are model accuracy (QM and force fields) & sampling statistics (SM)

Model Accuracy:

- **QM needs lower cost-scaling algorithms with high accuracy, and parallelization implementations are still active research areas**
- **High accuracy empirical FFs suffer milder but similar problems**
- **Lower accuracy empirical FFs and CG Modeling can be appropriate depending on science (much larger systems, timescales)**

Sampling Statistics:

- **Sampling to convergence is starting to become a more dominant issue (models are much better!): Replica exchange is not a panacea!**
- **Interesting (viable?) hardware: DEShaw, Folding@home, GPUs**

Potential Impact of HEC in Bio/Chemical Sciences

Given outlined "Grand Challenges" of 4 science areas, what is needed to "solve" Grand Challenges is High-End Capability Computing

- **Mathematical Models (especially in QM domain)**
- **Sampling Algorithms (especially in SM domain)**
- **Parallelization Software Support**
- **Traditional Facilities (Hardware, compilers, libraries (I/O) , mass storage; visualization and analysis tools; access)**
- **Speciality hardware (Anton (De Shaw); GPUs)**
- **Training/Education (important!)**

Community Codes/Software

Comp. Biophysicists do not work as teams! But consortiums develop to generate community codes/models

SM (MD): AMBER, TINKER, NAMD, GROMACS, LAMMPS, CHARMM, DLPOLY, ilmm, in-house codes

QM: GAMESS, Gaussian03, Q-Chem

Multiscale: PARASIM/LAMMPS/DYNAMO

Analysis: Gnuplot, Mathematica, Matlab, PERL, VMD

Memory requirements/core: 1-2GB
Core Hours/yr: 250K-30.0M (but multiply by ~10 users)
Typically bundle jobs to get to ~2.5-30K cores
Wall Clock: limited by queue times; w/restarts can be ~1-6 months
Online Storage: 50GB-2.5TB (accessible during simulation)
Archival storage for large datasets less critical?

AMBER9.0/10.0

Written in Fortran77/90 and uses MPI on most basic applications. Uses PME that scales as O(NlogN), and parallelized with MPI

Hybrid MPI/OpenMP implementation is available (but experimental) in AMBER10.0. Optimize cache

However, optimized parallel version works with only some subset of theoretical models.

Overlayed on top of fine-grained parallelization (if it exists for a given theoretical model) is another layer of (trivial) coarse-grained parallelization involving the replica exchange sampling algorithm, which runs N-independent simulations (each at a different temperatures), that involve infrequent communication to swap state point information (position and velocities of all atoms).

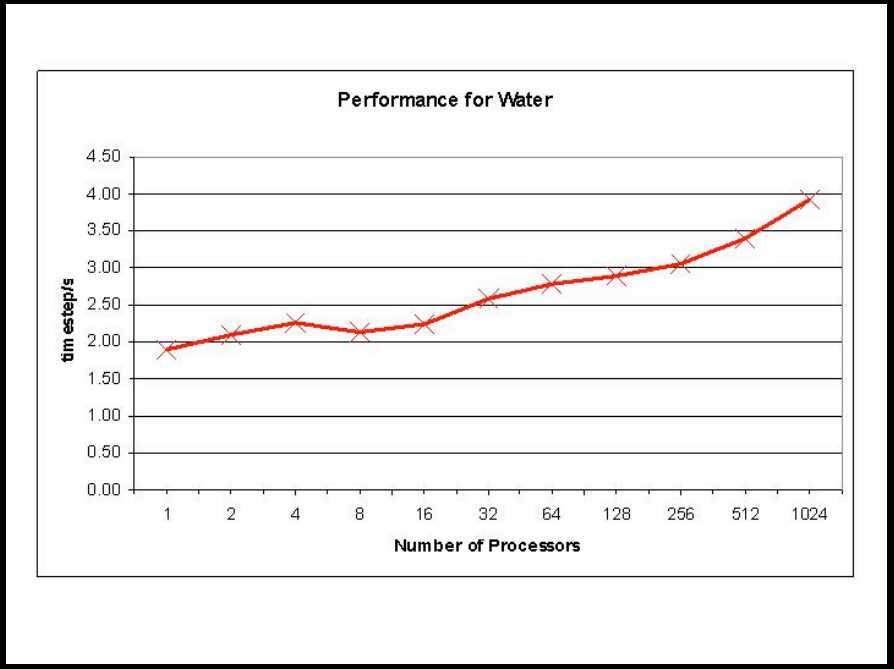
Strong Scaling

The PMEMD executable (MD kernal) of AMBER9.0 has been thoroughly tested by the author, Bob Duke of UNC Chapel Hill, on NERSC's Bassi and Jacquard machines, as well as the Cray-XT3 machines. Also worked on with NERSC's David Skinner to port and improve the performance of PMEMD (part of NERSC's benchmark suite and SC05 research activities).

Platform	*No. Proc.*	*Nsec/ Day*	*Scale (%)*	*Platform*	*No. Proc.*	*Nsec/ day*	*Scale (%)*	*Platform*	*No. Proc.*	*Nsec/ day*	*Scale (%)*
IBM	*4*	*0.48*	*98*	*Cray*	*4*	*0.40*	*92*	*Opteron*	*4*	*0.44*	*97*
SP5	*16*	*1.78*	*92*	*XT3*	*16*	*1.56*	*89*	*Infiniband*	*16*	*1.58*	*87*
(Bassi)	*32*	*3.26*	*84*	*(TBA)*	*32*	*2.80*	*80*	*Cluster*	*32*	*2.77*	*76*
	64	*6.06*	*78*		*64*	*4.70*	*67*	*(Jacquard)*	*64*	*4.91*	*68*
	128	*9.74*	*63*		*128*	*7.69*	*55*		*128*	*7.04*	*49*
	256	*12.23*	*39*		*256*	*9.60*	*34*		*256*		
	320	*13.50*	*35*		*320*	*9.89*	*28*		*320*		

Weak Scaling

Time increases from 1.9s on 1 processor for 21K particles, to 3.9s on 1024 processors for 21M particles. Both ewald terms must be calculated O(NlogN) as well as constraint forces- although latter are short ranged and should scale as O(N), their calculation requires a large number of short messages to be sent, and some latency effects become appreciable.



www.cse.scitech.ac.uk/arc/dlpoly_scale.shtml

Community Needs over Next 5 Years

Weak scaling and Strong scaling pretty good for MD. Thus concurrency and memory trends okay (assuming underlying libraries are keeping apace); Concurrency and memory trends not okay for QM.

Hardware: Low latency networks, would like generous cache at all levels.

Service issues: queue sizes and time limits requires lots of babysitting. Stability and turn-around is overall rate limiting and will likely continue to be so.

Analysis of large data sets: this is starting to become non-trivial