Dynameomics: Protein Mechanics, Folding and Unfolding through Large Scale All-Atom Molecular Dynamics Simulations

INCITE 6
David A. C. Beck
Valerie Daggett Research Group

Department of Medicinal Chemistry
University of Washington, Seattle
November 15th, 2005
Proteins

• Proteins are life’s machines, tools and structures
  – Many jobs, many shapes, many sizes
Proteins

• Proteins are life’s machines, tools and structures
  – Nature reuses designs for similar jobs
Proteins

- Proteins are hetero-polymers of specific sequence
  - There are 20 common polymeric units (amino acids)
    - Composed of a variety of basic chemical moieties
  - Chain lengths range from 40 amino acids on up
Proteins

- Proteins are hetero-polymers that adopt a unique fold
• Protein folding as a reaction
Proteins

- Protein folding ...

Transition state

Free Energy

Native

Denatured / Partially Unfolded

Good

Bad
Proteins

- Folded proteins

![Graph showing the free energy landscape of proteins with states: Bad, Native, and Good. The Native state is the lowest energy state and is the biologically relevant state.](image-url)
Proteins

- Folded proteins

Static, 3D coordinates of some proteins’ atoms are available from x-ray crystallography & NMR
Proteins

- Folded proteins

Static, 3D coordinates of some proteins’ atoms are available from PDB
http://www.pdb.org
Proteins

- Folded proteins are complex and dynamic molecules
Proteins

- Folded proteins are complex and dynamic molecules
Molecular Dynamics

• MD provides atomic resolution of native dynamics

PDB ID: 3chy, *E. coli* CheY 1.66 Å X-ray crystallography
Molecular Dynamics

- MD provides atomic resolution of native dynamics

PDB ID: 3chy, E. coli CheY 1.66 Å X-ray crystallography
Molecular Dynamics

- MD provides atomic resolution of native dynamics
Molecular Dynamics

- MD provides atomic resolution of native dynamics

3chy, waters added (i.e. solvated)
Molecular Dynamics

- MD provides atomic resolution of native dynamics

3chy, waters and hydrogens hidden
Molecular Dynamics

- MD provides atomic resolution of native dynamics

native state simulation of 3chy at 298 Kelvin, waters and hydrogens hidden
Proteins

- Folding & unfolding at atomic resolution

Free Energy

Bad

Good

Denatured / Partially Unfolded

Disordered, non-functional, heterogeneous ensemble of conformers
Proteins

• Protein folding, why we care how it happens

Many diseases are related to protein folding and/or misfolding in response to genetic mutation.
Proteins

- Protein folding, why we care how it happens

Transition state

Free Energy

Mutation

Native

Denatured / Partially Unfolded

Mutation

We need to comprehend folding to build nano-scale biomachines (that could produce energy, etc...)
Proteins

- Protein folding takes > 10 µs (often much longer)

[Diagram showing free energy landscape with states such as Native, Transition state, Denatured/Partially Unfolded, and Free Energy scales from Good to Bad]
Proteins

• Protein folding is the reverse of protein unfolding
Proteins

• Protein unfolding is relatively invariant to temperature

Free Energy

Native

Transition state

Denatured /
Partially Unfolded

Temperature

Good

Bad
Molecular Dynamics

- MD provides atomic resolution of folding / unfolding

unfolding simulation (reversed) of 3chy at 498 Kelvin, waters & hydrogens hidden
Molecular Dynamics

• Classically evolves an atomic system with time
  – Potential function (a.k.a force field)
    • Describes the energies of interaction between atom centers
  – Integration algorithm
    • Time dependent evolution of atomic coordinates in response to potential energy
  – Statistical sampling ensemble
    • Fixed thermodynamic variables, i.e. NVE
    • Number of atoms, box Volume, total Energy

Molecular Dynamics

- Potential function for MD$^{1,2}$

$$U = \text{Bond} + \text{Angle} + \text{Dihedral} + \text{van der Waals} + \text{Electrostatic}$$

Molecular Dynamics

• Potential function for MD$^{1,2}$

\[ U = \text{Bond} + \text{Angle} + \text{Dihedral} + \text{van der Waals} + \text{Electrostatic} \]

Molecular Dynamics

- Potential function for MD$^{1,2}$

$$U = \text{Bond} + \text{Angle} + \text{Dihedral} + \text{van der Waals} + \text{Electrostatic}$$

Molecular Dynamics

• Potential function for MD\textsuperscript{1,2}

\[ U = \text{Bond} + \text{Angle} + \text{Dihedral} + \text{van der Waals} + \text{Electrostatic} \]

Molecular Dynamics

- Potential function for MD\(^1,2\)

\[ U = \text{Bond} + \text{Angle} + \text{Dihedral} + \text{van der Waals} + \text{Electrostatic} \]

Molecular Dynamics

- Potential function for MD\(^1,2\)

\[ U = \text{Bond} + \text{Angle} + \text{Dihedral} + \text{van der Waals} + \text{Electrostatic} \]

Molecular Dynamics

- Non-bonded components of potential function

\[ U_{nb} = \text{van der Waals} + \text{Electrostatic} \]

\[ \sum_{\text{pairs } i,j} \left[ \varepsilon_{ij} \left( \frac{r_{0,ij}}{r_{ij}} \right)^{12} - 2\varepsilon_{ij} \left( \frac{r_{0,ij}}{r_{ij}} \right)^6 \right] \]

- To a large degree, protein structure is dependent on non-bonded atomic interactions
Molecular Dynamics

- Non-bonded components of potential function

\[ U_{nb} = \text{van der Waals} + \text{Electrostatic} \]
Molecular Dynamics

- Non-bonded components of potential function

\[ U_{nb} = \text{van der Waals} + \text{Electrostatic} \]
Molecular Dynamics

• Non-bonded components of potential function

\[ U_{nb} = \text{van der Waals} + \text{Electrostatic} \]
Molecular Dynamics

• Non-bonded components of potential function

\[ U_{nb} = \text{van der Waals} + \text{Electrostatic} \]
Molecular Dynamics

- Non-bonded components of potential function

\[ U_{nb} = \text{van der Waals} + \text{Electrostatic} \]

**NOTE:**
Sum over all pairs of N atoms, or
\[
\frac{N \times N - 1}{2} \text{ pairs}
\]

N is often between \(5 \times 10^5\) to \(5 \times 10^6\)

For \(5 \times 10^5\) that is \(1.25 \times 10^{11}\) pairs

THAT IS A LOT OF POSSIBLE PAIRS!
Molecular Dynamics

- Time dependent integration of classical equations of motion

\[ F = -\frac{\partial U}{\partial x} \]
\[ F = ma \]
\[ a = \frac{v_2 - v_1}{\partial t} \]
\[ v = \frac{x_2 - x_1}{\partial t} \]
\[ E = U + K \]
\[ \partial t = 2 \text{ fs} \]
Molecular Dynamics

- Time dependent integration

\[ F = -\frac{\partial U}{\partial x} \]

\[ F = ma \]

\[ a = \frac{v_2 - v_1}{\partial t} \]

\[ v = \frac{x_2 - x_1}{\partial t} \]

\[ E = U + K \]

\[ \partial t = 2 \text{ fs} \]
Molecular Dynamics

- Time dependent integration

\[ F = -\frac{\partial U}{\partial x} \]
\[ F = ma \]
\[ a = \frac{v_2 - v_1}{\partial t} \]
\[ v = \frac{x_2 - x_1}{\partial t} \]
\[ E = U + K \]
\[ \partial t = 2 \text{ fs} \]
Molecular Dynamics

- Time dependent integration

\[ F = -\frac{\partial U}{\partial x} \]
\[ F = mg \]
\[ a = \frac{v_2 - v_1}{\partial t} \]
\[ v = \frac{x_2 - x_1}{\partial t} \]
\[ E = U + K \]
\[ \partial t = 2 \text{ fs} \]
Molecular Dynamics

- Time dependent integration

\[
F = -\frac{\partial U}{\partial x}
\]

\[
F = ma
\]

\[
a = \frac{v_2 - v_1}{\partial t}
\]

\[
v = \frac{x_2 - x_1}{\partial t}
\]

\[
E = U + K
\]

\[
\partial t = 2 \text{ fs}
\]
Molecular Dynamics

- Time dependent integration

\[ F = -\frac{\partial U}{\partial x} \]
\[ F = ma \]
\[ a = \frac{v_2 - v_1}{\partial t} \]
\[ v = \frac{x_2 - x_1}{\partial t} \]
\[ E = U + K \]
\[ \partial t = 2 \text{ fs} \]
Molecular Dynamics

- Time dependent integration

\[ F = -\frac{\partial U}{\partial x} \]

\[ F = ma \]

\[ a = \frac{v_3 - v_2}{\partial t} \]

\[ v = \frac{x_3 - x_2}{\partial t} \]

\[ E = U + K \]

\[ \partial t = 2 \text{ fs} \]
Molecular Dynamics

- Time dependent integration

Evaluate forces and perform integration for every atom

Each picosecond of simulation time requires 500 iterations of cycle

E.g. with 50,000 atoms, each ps ($10^{-12}$ s) involves 25,000,000 evaluations

\[
F = -\frac{\partial U}{\partial x}
\]
\[
F = ma
\]
\[
a = \frac{v_3 - v_2}{\partial t}
\]
\[
v = \frac{x_3 - x_2}{\partial t}
\]
\[
E = U + K
\]
\[
\partial t = 2 \text{ fs}
\]
Molecular Dynamics

- Scalable, parallel MD & analysis software:

\[
\text{ilmm}
\]

*in lucem* Molecular Mechanics\(^1\)

Molecular Dynamics

• *ilmm* is written in C (ANSI / POSIX)
• 64 bit math
• POSIX threads / MPI

**Software design philosophy:**
- **Kernel**
  - Compiles user’s molecular mechanics programs
  - Schedules execution across processor and machines
- **Modules, e.g.**
  - Molecular Dynamics
  - Analysis

**Diagram:**
- POSIX threads (multiprocessor machines)
- Message Passing Interface (multiple machines)
  - **VERY** high bandwidth
Molecular Dynamics

- *ilmm* is written in C (ANSI / POSIX)
- 64 bit math
- POSIX threads / MPI

**Software design philosophy:**
- Kernel
  - Compiles user’s molecular mechanics programs
  - Schedules execution across processor and machines
- Modules, e.g.
  - Molecular Dynamics
  - Analysis

POSIX threads (multiprocessor machines) + Message Passing Interface (multiple machines)

*VERY* high bandwidth
Dynameomics

- Simulate representative protein from all folds
Dynameomics

- Simulate representative protein from all folds
  - Nature reuses designs for similar jobs
Dynameomics

- Simulate representative protein from all folds

150 folds represent ~ 75% of known protein structures

Dynameomics

• Simulate representative protein from all folds
  – Native (folded) dynamics
    • 20 nanosecond simulation at 298 Kelvin
  – Folding / unfolding pathway
    • 3 x 2 ns simulations at 498 K
    • 2 x 20 ns simulations at 498 K
  – Each target requires 6 simulations

= MANY CPU HOURS
Dynameomics

- NERSC DOE INCITE award
  - 2,000,000 + hours
  - 906 simulations of 151 protein folds on Seaborg
    - One to two simulations per node (8 – 16 CPUs / simulation)
    - Opportunity to tune ilmm for maximum performance
Dynameomics

- Load balancing
  - Even distribution of non-bonded pairs to processors

5 minute (wall clock) run
16 threads on 16 CPU IBM Nighthawk

- ~20% faster
• **Parallel efficiency**
  
  – **Threaded computations on 16 CPU IBM Nighthawk**

Parallel efficiency, $e(p) = \left( \frac{1}{p} \right) \left( \frac{t(1)}{t(p)} \right)$

$p$, number of processors

$t(p)$, run-time using $p$ processors
Dynameomics

• Simulate representative from top 151 folds
  – 151 folds represent about 75% of known proteins
  • ~ 11 µs of combined sim. time from 906 sims!
  • ~ 2 terabytes of data (w/ 40 to 60% compression!)
  • ~ 75 / 151 have been analyzed
  • Validated against experiment where possible
Dynameomics

• Now what?
  – Simulate the top 1130 folds (>90%)
    • More CPU time
  – Share simulation data from top 151 folds w/ world:
    • Coordinates, analyses, available via WWW
    • MicrosoftSQL database w/ On-Line Analytical Processing (OLAP)
      • End-user queries of coordinate data, analyses, etc.
  – Data mining
    • More CPU time, clever statistical algorithms, etc.

www.dynameomics.org
Acknowledgements

• DOE / NERSC’s INCITE  (David Skinner, et al)
• NIH
• Microsoft, Inc.
• Structures rendered using Chimera, Molscript, Raster3D & PyMOL