Atomic Simulations of Nanoscale Molecular Motion

Zhiyong Zhang, Willy Wriggers
School of Health Information Sciences & Institute of Molecular Medicine
University of Texas – Houston
Protein Dynamics is Hierarchical

Vibration of bonds: $10^{-15}$ s

Protein folding/unfolding

$10^{-6}$ s, $10^{-3}$ s, s or even longer

Large-scale functional motions
From experiments to theory

Experimental techniques:

* X-ray crystallography, NMR, Cryo-EM etc

Computer Simulations
Molecular Dynamics Simulation

\[ \frac{d^2 r_i}{dt^2} = \frac{F_i(r_1, r_2, \cdots, r_n)}{m_i} \]

\[ F_i(r_1, r_2, \cdots, r_n) = -\nabla V(r_1, r_2, \cdots, r_n) \quad i = 1, 2, \cdots, N \]

\[ V_i(r) = V_i(r_1, r_2, r_3, \cdots, r_N) \]

\[ = \sum_{\text{bonds}} \frac{1}{2} K_b (b - b_0)^2 + \sum_{\text{angles}} \frac{1}{2} K_q (q - q_0)^2 + \sum_{\text{improper}} \frac{1}{2} K_x (x - x_0)^2 + \]

\[ \sum_{\text{dihedral}} K_j \left[ 1 + \cos(n_j - d) \right] + \sum_{ij} \left[ \frac{C_{12}}{r_{ij}^{12}} - \frac{C_6}{r_{ij}^6} - \frac{q_i q_j}{4\pi \varepsilon_0 \varepsilon_r r_{ij}} \right] \]

NAMD, Amber, CHARMM, Gromos, etc.
Applications of MD

- Conformational space search
- Equilibrium state of the system
- Actual protein dynamics

Sampling Problem

Rotation of buried sidechains
Local denaturations
Allosteric transitions

Hinge bending
Rotation of surface sidechains
Elastic vibrations

Bond stretching
Molecular dynamics timestep

Energy Surface →
Exploration by Simulation.

© Jeremy Smith
Sampling techniques

• Umbrella sampling
• Targeted molecular dynamics
• Steered molecular dynamics
• Methods based on collective coordinates
A Simple Example of Collective coordinate

\[ z = \frac{\sqrt{2}}{2}x + \frac{\sqrt{2}}{2}y \]
Collective coordinates in proteins

- Diagonalize Hessian matrix

\[ C = U \Lambda U^T \]

- Principal Component Analysis

\[ C_{ij} = \left\langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \right\rangle \]

- Normal Mode Analysis

\[ C_{ij} = \frac{\partial^2 V}{\partial x_i \partial x_j} \]
Eigenvalues and Eigenvectors

Matrix algebra

Online introduction, e.g.

http://www.sosmath.com/matrix/matrix.html
Principal Component Analysis

Can be applied to MD simulation trajectories to detect the global, correlated motions of the system (the principal components).

Why are the PCs important?

Amadei et al. argue that we can separate the configurational space into 2 sub-spaces:

1. The Essential subspace: correlated motions comprising only a few of the degrees of freedom available to the protein = FUNCTIONALLY IMPORTANT

2. The “Irrelevant” subspace: independent, Gaussian fluctuations, which are constrained and of no/little functional relevance – act locally
Overview

MD trajectory

Construct & diagonalize covariance matrix

Eigenvectors, ranked by eigenvalue

Analyse eigenvectors: eg. Visualisation of motions

First 2 eigenvectors account for 60% of total positional fluctuations

Global, concerted motions

Constrained, fluctuations

© Sansom lab, http://indigo1.biop.ox.ac.uk/MD_workshops/
Visualizing PCs

Porcupine plots can be used to display the motion described by an eigenvector in a static image.

A cone extending from the C-alpha position shows the direction of the atom along the eigenvector.

Covariance plots are a tool to visualize atoms which have a high correlation coefficient from the covariance matrix.

© http://dynamite.biop.ox.ac.uk/dynamite
Sampling techniques based on collective coordinates

*drive MD by collective coordinates (PCA or NMA)*

First approach with PCA: “Essential Molecular Dynamics”


Use the PCs from free MD to drive a protein from one conformation to another. Used by Daidone et al. to study Cytochrome c folding with MD. Only 106 degrees of freedom out of a total 3000 were used to bias the simulation

- Conformational Flooding and Chemical Flooding
- Amplified Collective Motions (ACM)
Anisotropic Network Model: ANM

Protein is equivalent to a three dimensional elastic network

\[
V = \sum_{i,j \neq i} \frac{1}{2} k_{ij} (r_{ij} - r_{ij}^0)^2
\]

\[
\Gamma_{ij} = \begin{cases} 
-k_{ij} \frac{(r_i^\alpha - r_j^\alpha)(r_i^\beta - r_j^\beta)}{r_{ij}^2} & |r_{ij} = r_{ij}^0 \quad i \neq j; \\
\sum_{j} k_{ij} \frac{(r_i^\alpha - r_j^\alpha)(r_i^\beta - r_j^\beta)}{r_{ij}^2} & |r_{ij} = r_{ij}^0 \quad i = j \end{cases}
\]

Weak-coupling method

\[ E_k(t) = \sum_{i=1}^{N} \frac{1}{2} m_i \vec{v}_i(t) \]  \hspace{1cm} \text{Kinetic energy}

\[ T(t) = \frac{2E_k(t)}{3Nk_B} \]  \hspace{1cm} \text{Actual temperature}

Temperature-scaling factor \( S = \left[ 1 + \frac{\Delta t}{\tau_T} \left[ \frac{T_0}{T(t)} - 1 \right] \right]^{1/2} \)

\[ \text{Used to scale velocities} \]
Amplified-collective-motion technique

\[ \{ \vec{V}_i ; \ i = 1, 2, \ldots, N \} \quad \text{Total velocities} \]

\[ \{ V^\alpha_c ; \ \alpha = 1, 2, \ldots, N_r \} \quad \text{Velocities of C.O.M of residues} \]

\[ \vec{V}_1 = \sum_{l=1}^{n} \left[ \sum_{\alpha=1}^{N_r} (V^\alpha_c \cdot e^\alpha_l) e_l \right] \quad \text{Velocity projection onto low-frequency NMs} \]

\[ \vec{V} = S_h \vec{V}_1 + S_f \vec{V}_2 \]

\[ \vec{V}_2 = \vec{V} - \vec{V}_1 \quad \text{Other velocities} \]

\[ T_0 = 800\text{K} \]

\[ T_0 = 300\text{K} \]
Folding/Unfolding of S-Peptide Analog


ACM 30-ns 3-modes @ 358K + other-DOF @ 274K
Normal modes are updated every 10-25 time steps
Control simulation 30-ns all-DOF 274K
implicit water model: Generalized Born model
Secondary structures (by DSSP)
Folding/Unfolding of S-Peptide Analog
Domain motions in Bacteriophage T4 lysozyme

Closure mode

(178L vs 152L)

Twist mode

(174L vs 150L)
Projections onto the Functional Subspace

ACM: 3-ns
3 modes @ 800K
others @ 300K
Modes are updated every 100 time steps
Standard MD 3-ns
all @ 300K
SPC water model
Domain motions in T4L
The myosin cross-bridge is a molecular machine with communicating functional units. How can the small changes at the active site be amplified into the large conformational changes? How do mutations interfere its functional dynamics?
ACM vs. MD: Myosin
Myosin - structure comparison

MD simulation (1ns)@300K:
6 degrees and 12 Angstrom

ACM simulation (1ns) (3@800K + others@300K):
31 degrees and 51 Angstrom
Global Collective Coordinates: What are the Limitations?

In NMA, we do not know \textit{a priori} which is a functionally relevant mode, the first 12 low-frequency modes are probable candidates.

In PCA, the global modes don’t converge due to time limitations of the molecular dynamics simulation (sampling problem):


Goal: an alternative statistical theory that describe dynamic features locally and that does not suffer from the sampling and orthogonalization problems.

Some ideas come from image processing, like face recognition.
Local Feature Analysis (LFA)

LFA is to derive local topographic representations for any class of objects. Unlike the global eigenmodes, they give a description of objects in terms of statistically derived local features and their positions.

Is LFA applicable to protein localized dynamics?

Local Feature Analysis (LFA)
- Theory (I)

Covariance matrix from the MD simulation: \( C(i, j) \equiv \langle \Delta x_i \Delta x_j \rangle \equiv \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \)

**PCA:** \( C(i, j) = \sum_{r=1}^{3N} \Psi_r(i) \lambda_r \Psi_r(j) \)  →  **PCA output:** \( A_r = \sum_{i=1}^{3N} \Psi_r(i) \Delta x_i \equiv \sum_{j=1}^{3N} K_r(i) \Delta x_i \)

**General form for the LFA kernel:** \( K(i, j) = \sum_{r, s=1}^{n} \Psi_r(i) Q_{rs} \Psi_s(j) \)  →  \( K(i, j) = \sum_{r=1}^{n} \Psi_r(i) \frac{1}{\sqrt{\lambda_r}} \Psi_r(j) \)

**LFA output:** \( O(i) \equiv \sum_{j=1}^{3N} K(i, j) \Delta x_j \)  →  \( O(i) = \sum_{j=1}^{3N} \left( \sum_{r=1}^{n} \Psi_r(i) \frac{1}{\sqrt{\lambda_r}} \Psi_r(j) \right) \Delta x_j = \sum_{r=1}^{n} \frac{A_r}{\sqrt{\lambda_r}} \Psi_r(i) \)

**Residual correlation:** \( \langle O(i)O(j) \rangle = \sum_{r=1}^{n} \Psi_r(i) \Psi_r(j) \equiv P(i, j) \)
Local Feature Analysis (LFA)

- Theory (II)

We replaced the n global PCA modes with the full 3N LFA output functions. Therefore an additional dimensionality reduction step is required in the LFA output space. We approximate the entire 3N outputs with only a small subset of them that correspond to the strongest local features by taking advantage of the fact that neighboring outputs are highly correlated.

Reconstruct the outputs:

\[ O^{rec}(i) = \sum_{m=1}^{\|A\|} a_m(i)O(i_m) \]

Optimal linear prediction coefficients:

\[ a_m(i) = \sum_{i_l \in A} P(i, i_l)(P^{-1})_{lm} \]

Average reconstruction mean square error:

\[ E^{rec} = \langle \|O^{err}(i)\|^2 \rangle \equiv \langle \|O(i) - O^{rec}(i)\|^2 \rangle \]
Sparse Distributions in T4L

(a) The first 4 PCA modes were used to do LFA, n=4; (b) n=8, (c) n=12, and (d) n=15. (e) Root-mean-square fluctuations of C_alpha atoms in T4L.
Output Functions’ Correlations

(a) The first 4 PCA modes were used to do LFA, n=4; (b) n=8, (c) n=12, and (d) n=15.

\[
\langle \vec{O}_h \cdot \vec{O}_k \rangle = \sum_{d=1}^{3} \langle O(h_d)O(k_d) \rangle = \sum_{d=1}^{3} P(h_d, k_d)
\]
Local Dynamic Domains in T4L

(a) t=0 ns, (b) t=4.00 ns, and (c) t=8.25 ns.
Four local features with different colors
Compare with Experimental Results

(a) Projection of x-ray structures, and the MD simulation, (b) the most open structure (178L), and (c) the most closed structure (152L).
Convergence of PCA and LFA

Different time windows have almost the same local features.
Convergence of PCA and LFA

The intrinsic dynamics of local domains is more extensively sampled than that of globally coherent PCA modes.
Local Feature Analysis of Myosin

Twelve seed atoms

Twelve local dynamical domains
Outlook: Predicting Functional Motion

- It appears that PCA and NMA over-estimate the coherence of global motion across large biopolymers and create artifacts due to orthogonalization.

- LFA captures local dynamic features reproducibly and is less sensitive to the MD sampling problem.

- There is hope for MD simulations of million-atom systems if we perform a statistical analysis that emphasizes dynamic domains that are moving independently from each other.
Future work

• ACM is a non-equilibrium simulation, how to recover the Boltzmann distribution and calculate thermodynamics properties?

• Improve the sparsification algorithm, and investigate the potential uses of LFA for applications in prediction, sampling and classification of large-scale macromolecular structure and dynamics.
Resources and Further Reading

WWW:
http://www.sosmath.com/matrix/matrix.html
http://starship.python.net/crew/hinsen/MMTK
http://dynamite.biop.ox.ac.uk/dynamite

Papers:
L. I. Smith “A tutorial on Principal Component Analysis” (2002) e.g. at
Amadei, Linsen, Berendsen, Proteins (1993), 17:412-425
Acknowledgement

• biomachina.org
• Dr. Danny Sorensen (Rice University)
• USTC: Prof. Haiyan Liu, Prof. Yunyu Shi

This work was supported by grants from NIH (1R01GM62968), Human Frontier Science Program (RGP0026/2003), Alfred P. Sloan Foundation (BR-4297), and a training fellowship from the Keck Center Pharmacoinformatics Training Program of the Gulf Coast Consortia (NIH Grant No.1 R90 DK071505-01).