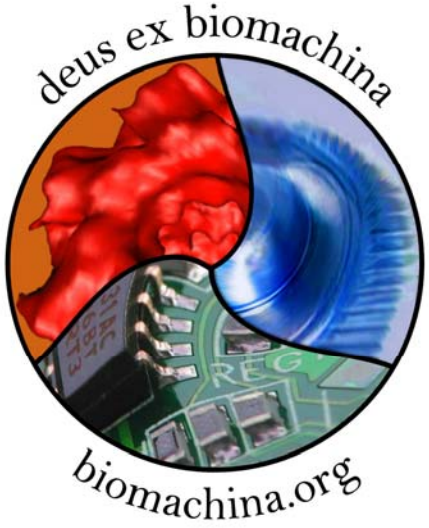
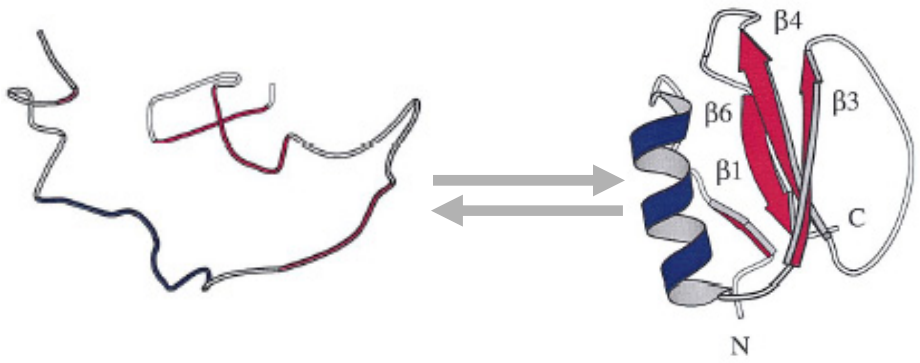


Atomic Simulations of Nanoscale Molecular Motion

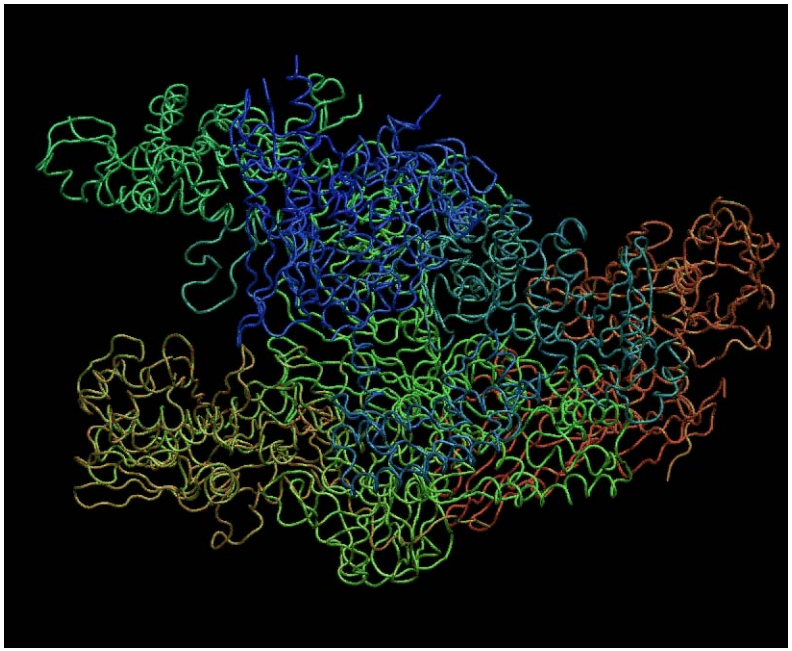
Zhiyong Zhang, Willy Wriggers
School of Health Information Sciences &
Institute of Molecular Medicine
University of Texas – Houston

Protein Dynamics is Hierarchical

Vibration of bonds: 10^{-15} s



Protein folding/unfolding



Large-scale functional motions

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

From experiments to theory

Experimental techniques:

X-ray crystallography, NMR, Cryo-EM etc

Computer Simulations

Molecular Dynamics Simulation

$$\frac{d^2 r_i}{dt^2} = F_i(r_1, r_2, \cdots, r_n) \Big/ m_i$$

$$F_i(r_1, r_2, \cdots, r_n) = -\nabla V(r_1, r_2, \cdots, r_n) \qquad i = 1, 2, \cdots, N$$

$$V_i(\vec{r}) = V_i(\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots, \vec{r}_N)$$

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

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

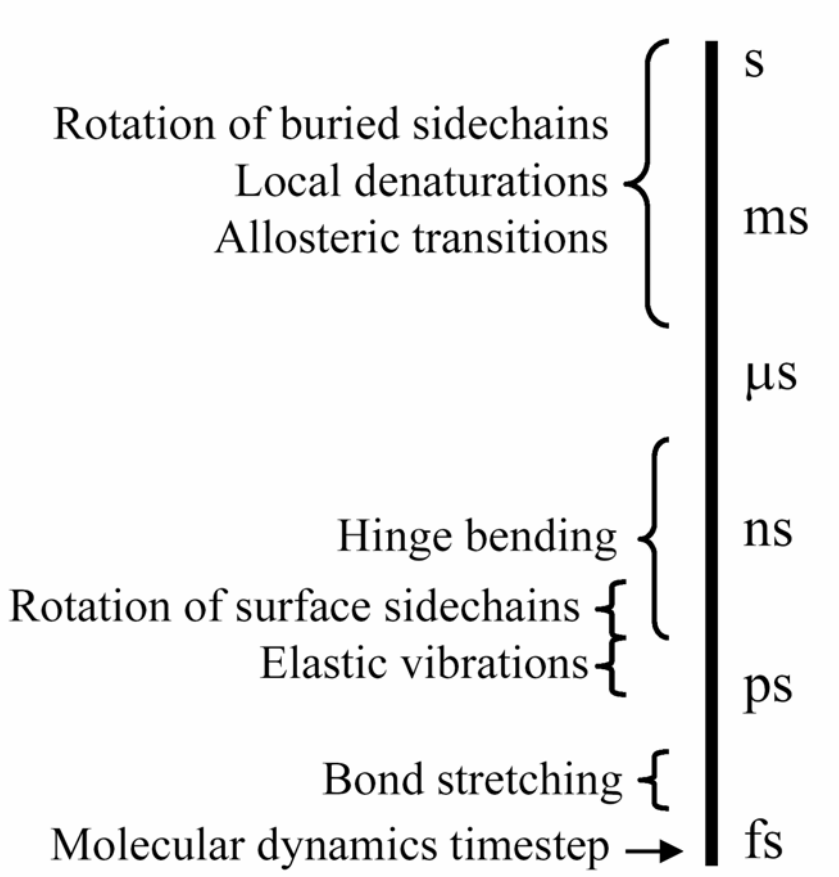
NAMD, Amber, CHARMM, Gromos, etc.

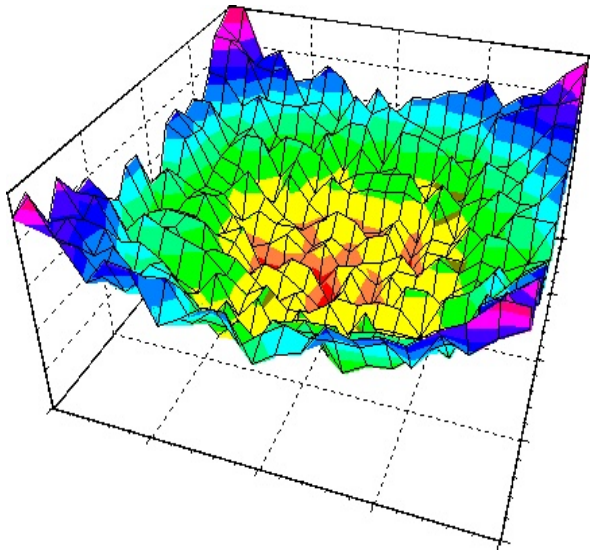
Applications of MD

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

Karplus M, McCammon JA: **Molecular dynamics simulations of biomolecules**. *Nature Struct Biol* 2002, **9:**646-652.

Sampling Problem



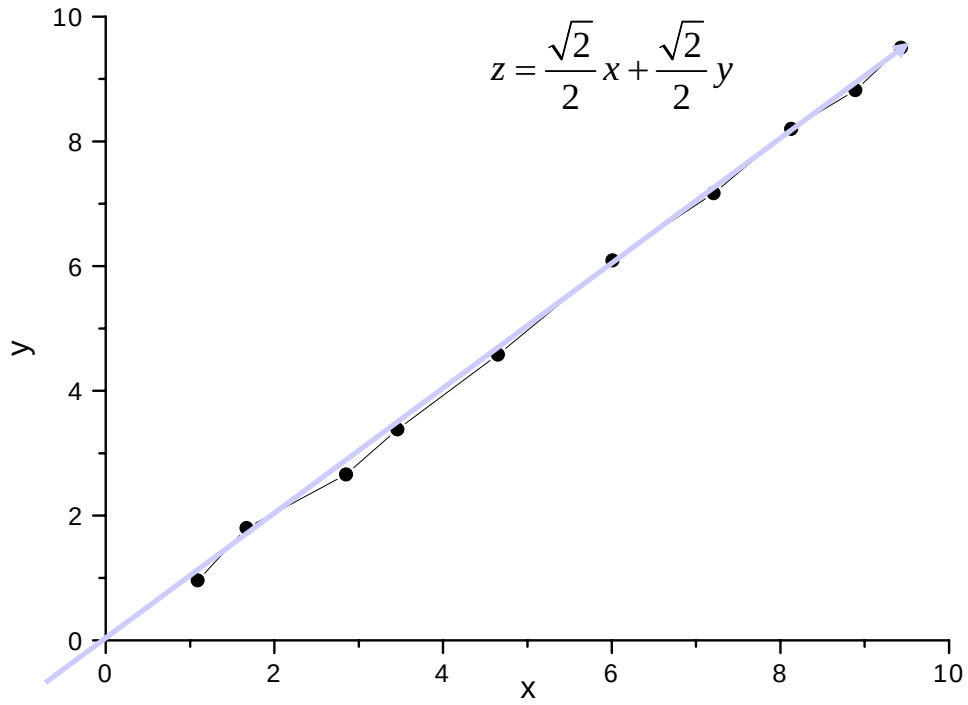


Energy Surface →
Exploration by Simulation.

Sampling techniques

- **Umbrella sampling**
- **Targeted molecular dynamics**
- **Steered molecular dynamics**
- **Methods based on collective coordinates**

A Simple Example of Collective coordinate



Collective coordinates in proteins

- **Diagonalize Hessian matrix**

$$C = U \Lambda U^T$$

- **Principal Component Analysis**

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

- **Normal Mode Analysis**

$$C_{ij} = \partial^2 V \big/ \partial x_i \, \partial x_j$$

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. **<u>The Essential subspace</u>: correlated motions comprising only a few of the degrees of freedom available to the protein = *FUNCTIONALLY IMPORTANT***

2. **<u>The "Irrelevant" subspace</u>: independent, Gaussian fluctuations, which are constrained and of no/little functional relevance – act locally**

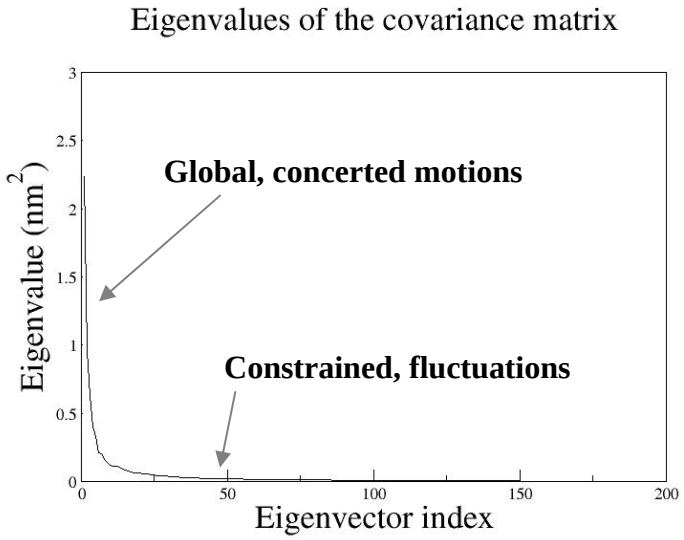
Overview

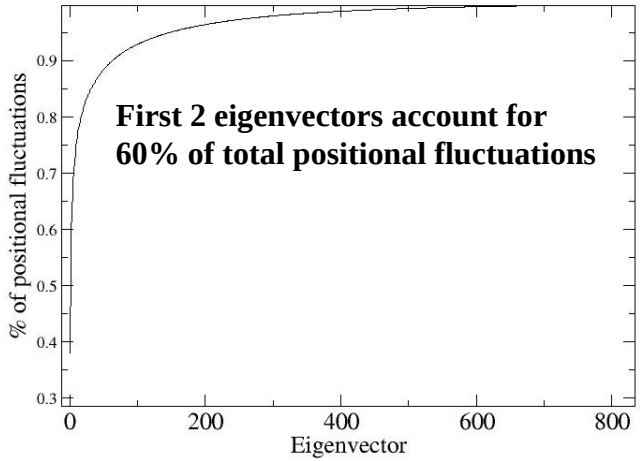
MD trajectory

Construct & diagonalize covariance matrix

Eigenvectors, ranked by eigenvalue

Analyse eigenvectors: eg. Visualisation of motions

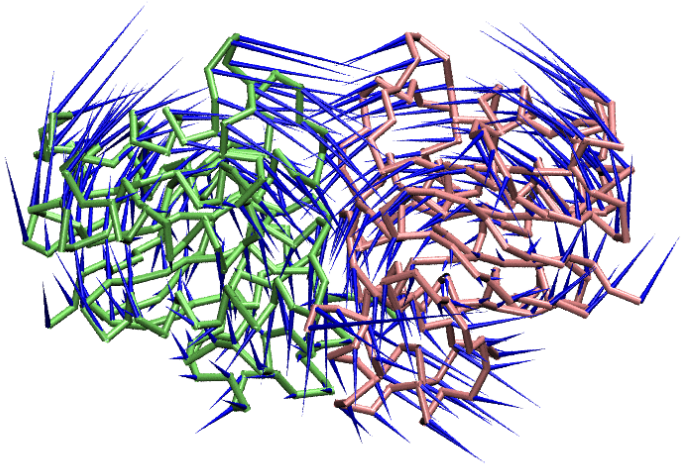




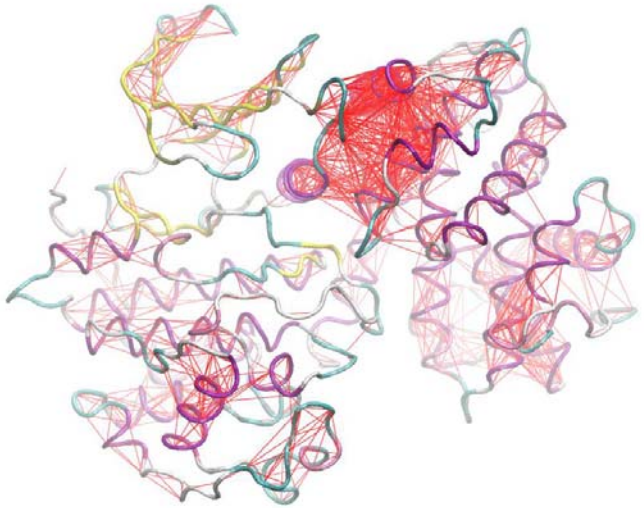
Visualizing PCs

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



Sampling techniques based on collective coordinates

drive MD by collective coordinates (PCA or NMA)

First approach with PCA: “Essential Molecular Dynamics”

Amadei, Linsen, Berendsen – Proteins (1993), 17:412-425

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

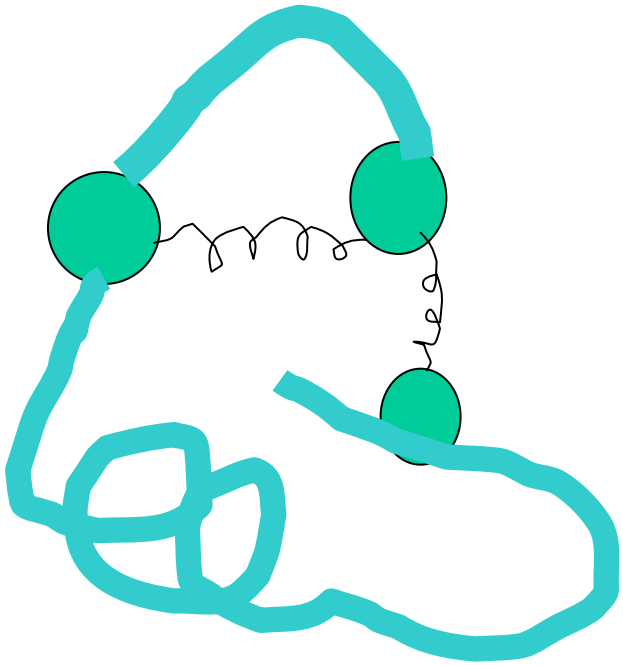
- **Conformational Flooding and Chemical Flooding**
- **Amplified Collective Motions (ACM)**

Anisotropic Network Model: ANM

Protein is equivalent to a three dimensional elastic network

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

$$\Gamma_{ij} = \begin{cases} -k_{ij} \dfrac{(r_i^\alpha - r_j^\alpha)(r_i^\beta - r_j^\beta)}{r_{ij}^2} \Big| r_{ij} = r_{ij}^0 & i \neq j; \\ \displaystyle\sum_j k_{ij} \dfrac{(r_i^\alpha - r_j^\alpha)(r_i^\beta - r_j^\beta)}{r_{ij}^2} \Big| r_{ij} = r_{ij}^0 & i = j \end{cases}$$



Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I: **Anisotropy of fluctuation dynamics of proteins with an elastic network model**. *Biophys J* 2001, **80:**505-515.

Weak-coupling method

$$E_k(t) = \sum_{i=1}^{N} \frac{1}{2} m_i \vec{V}_i^{\,2}(t)$$ **Kinetic energy**

↓

$$T(t) = {2E_k(t)}\Big/{3Nk_B}$$ **Actual temperature**

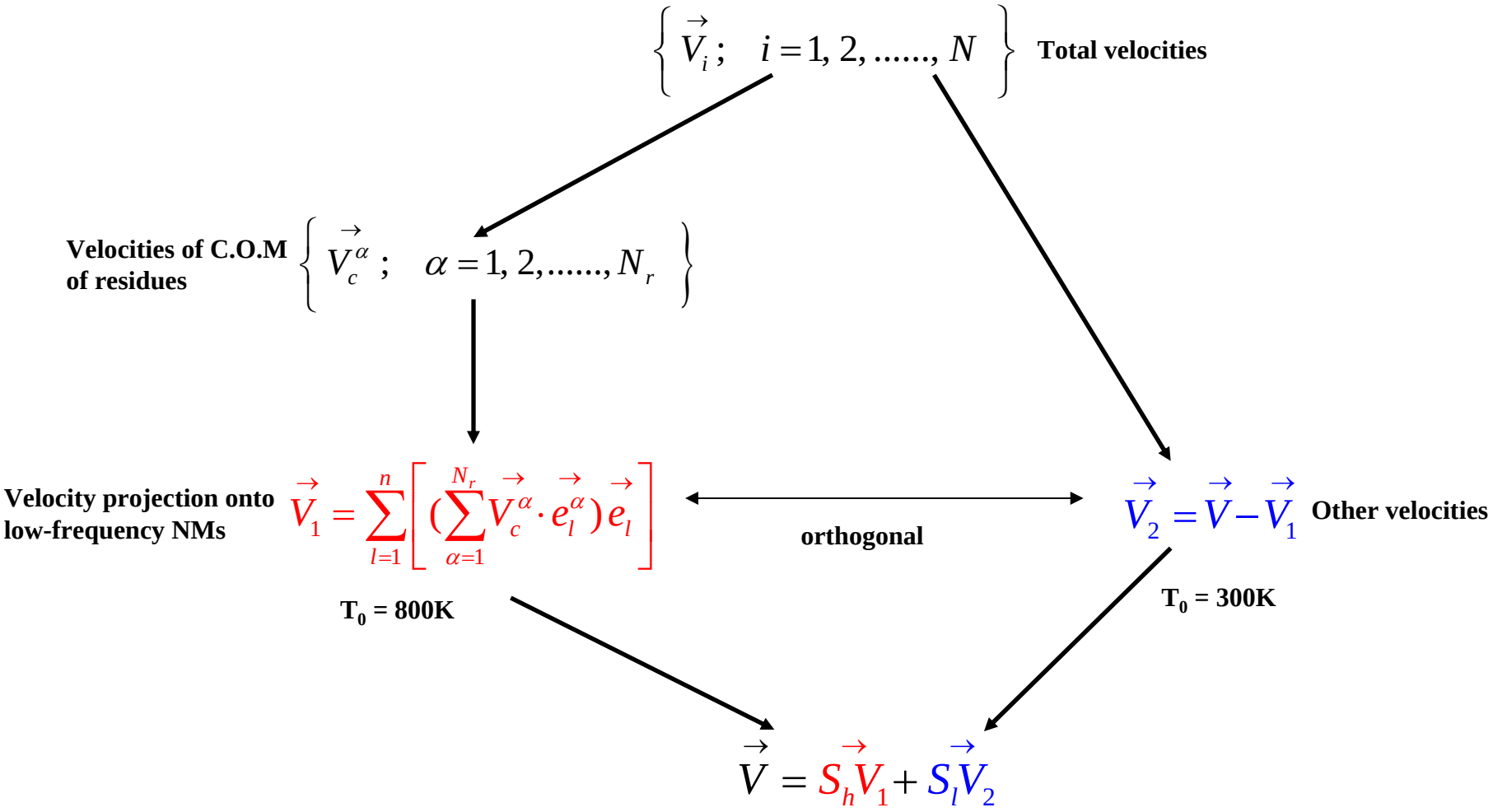
↓

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

↓

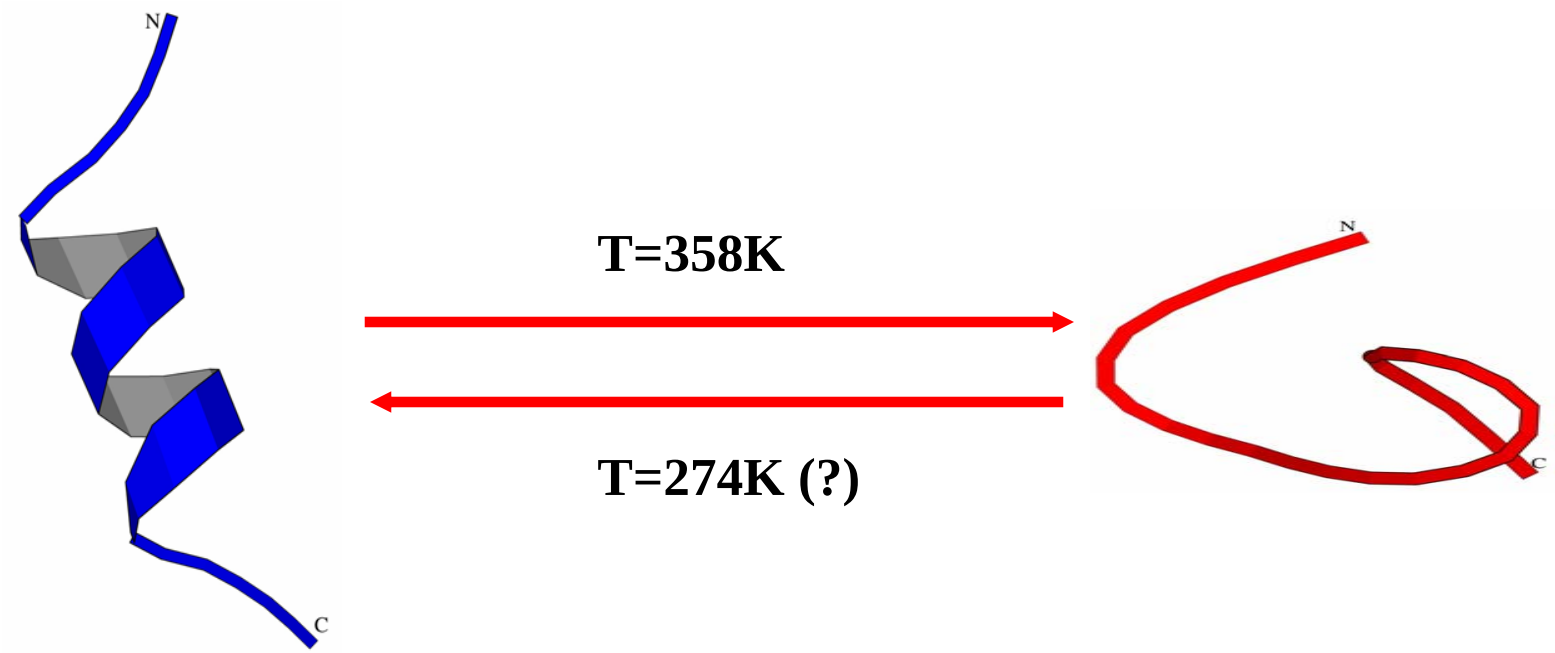
Used to scale velocities

Amplified-collective-motion technique



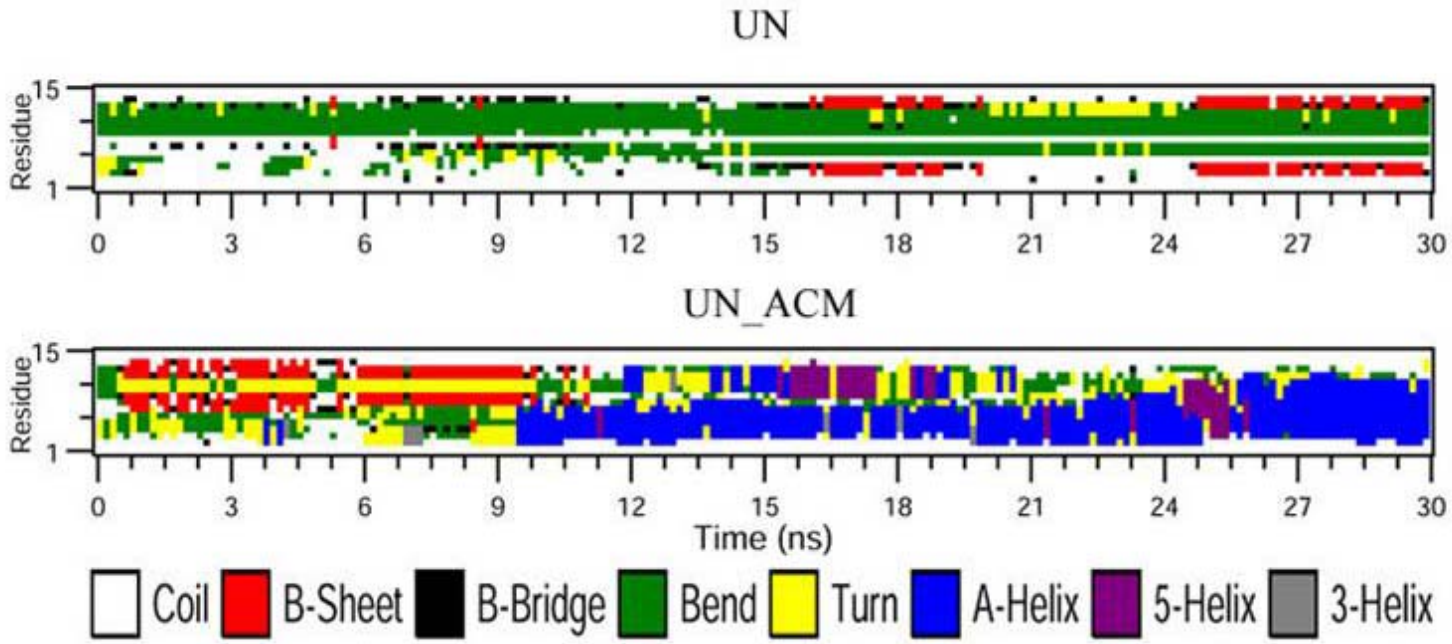
Folding/Unfolding of S-Peptide Analog

Zhang et al., Biophys J. (2003) 84:3583-93.



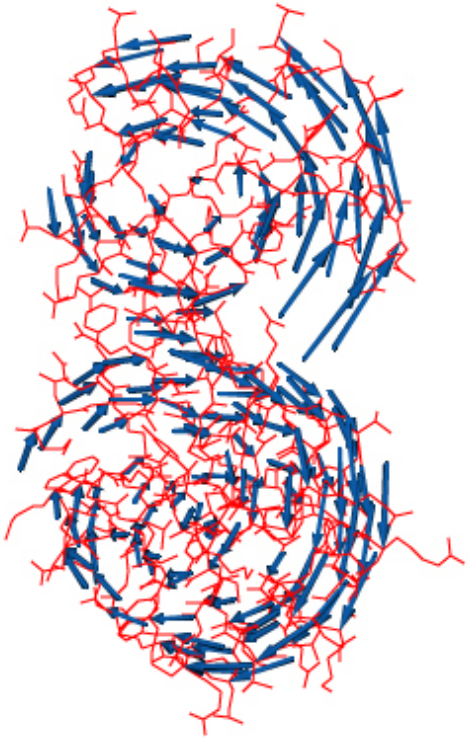
ACM 30-ns 3-modes @ 358K + other-DOF @ 274K
Normal modes are updated every 10-25 time steps
Control simulation 30-ns all-DOF 274K
implicit water model: Generalized Born model

Secondary structures (by DSSP)



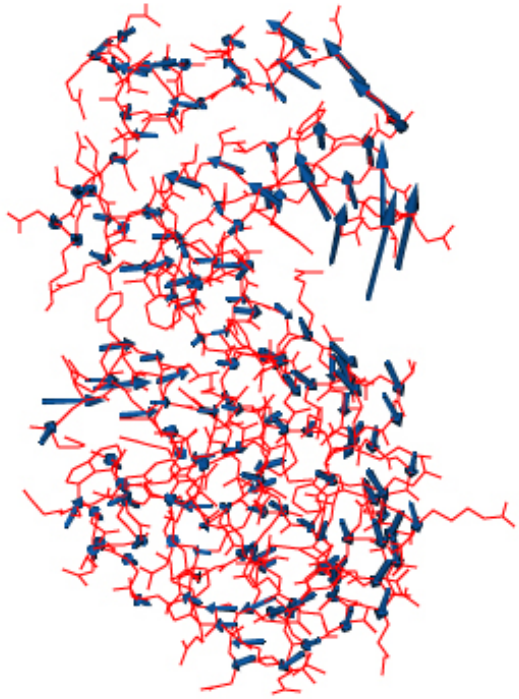
Folding/Unfolding of S-Peptide Analog

Domain motions in Bacteriophage T4 lysozyme



Closure mode

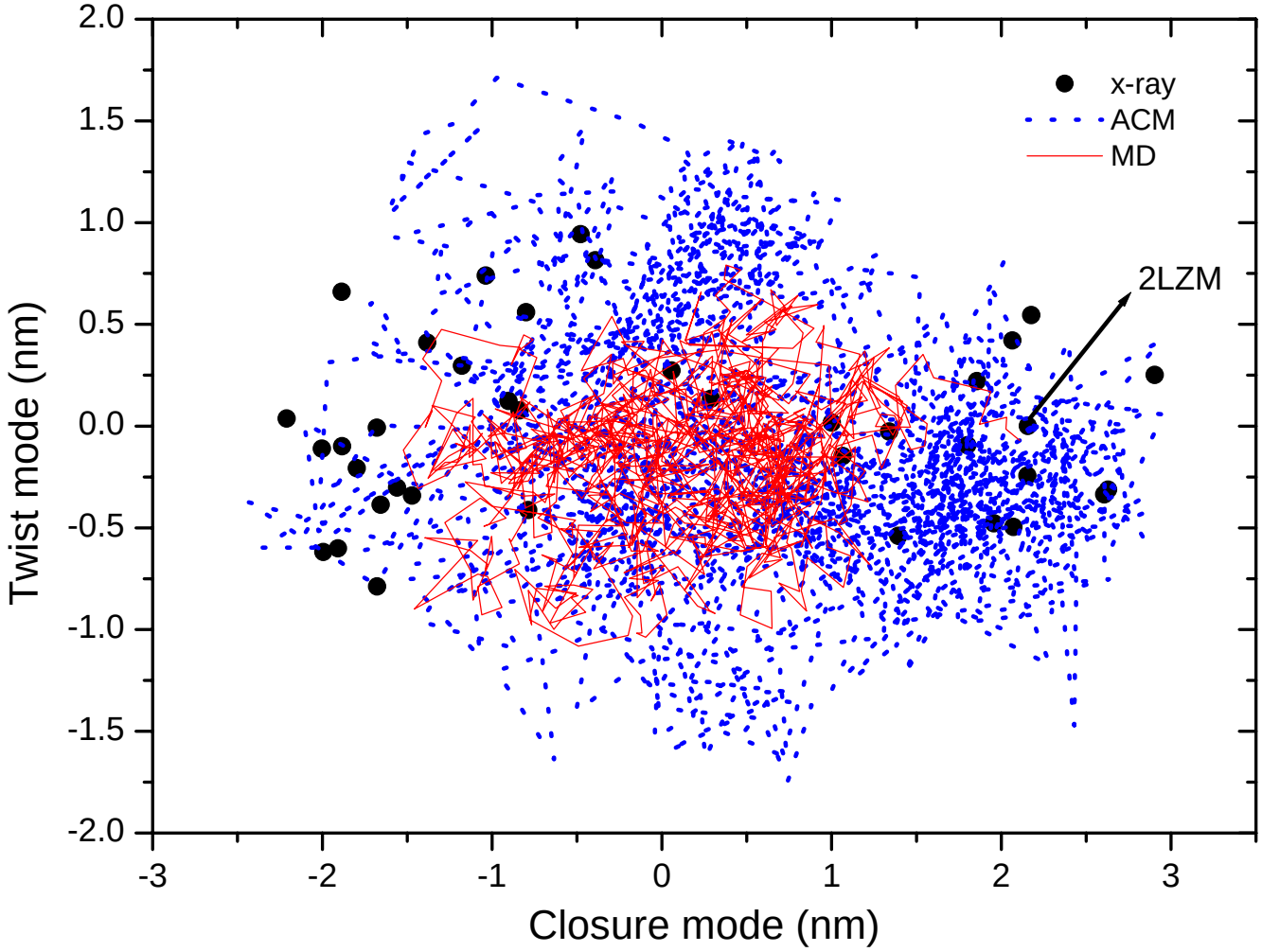
(178L vs 152L)



Twist mode

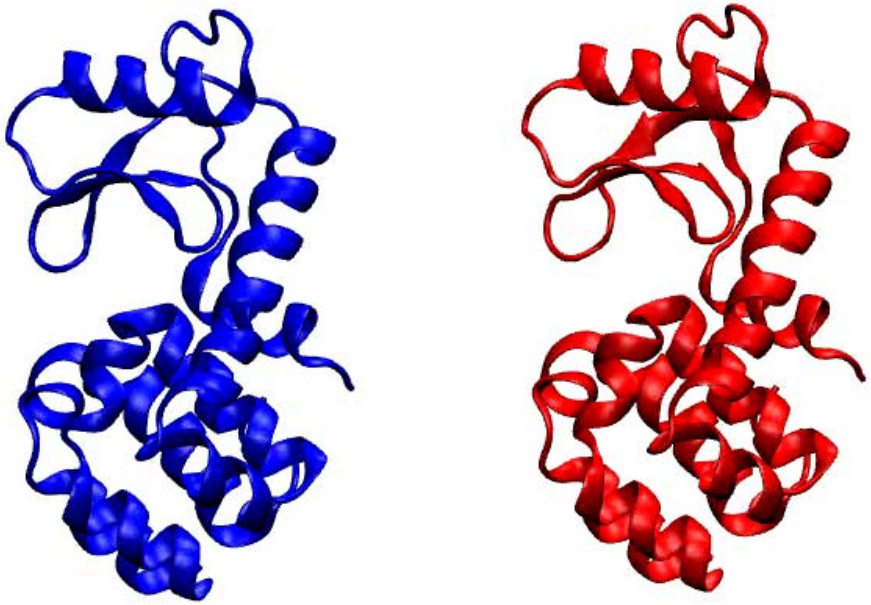
(174L vs 150L)

Projections onto the Functional Subspace

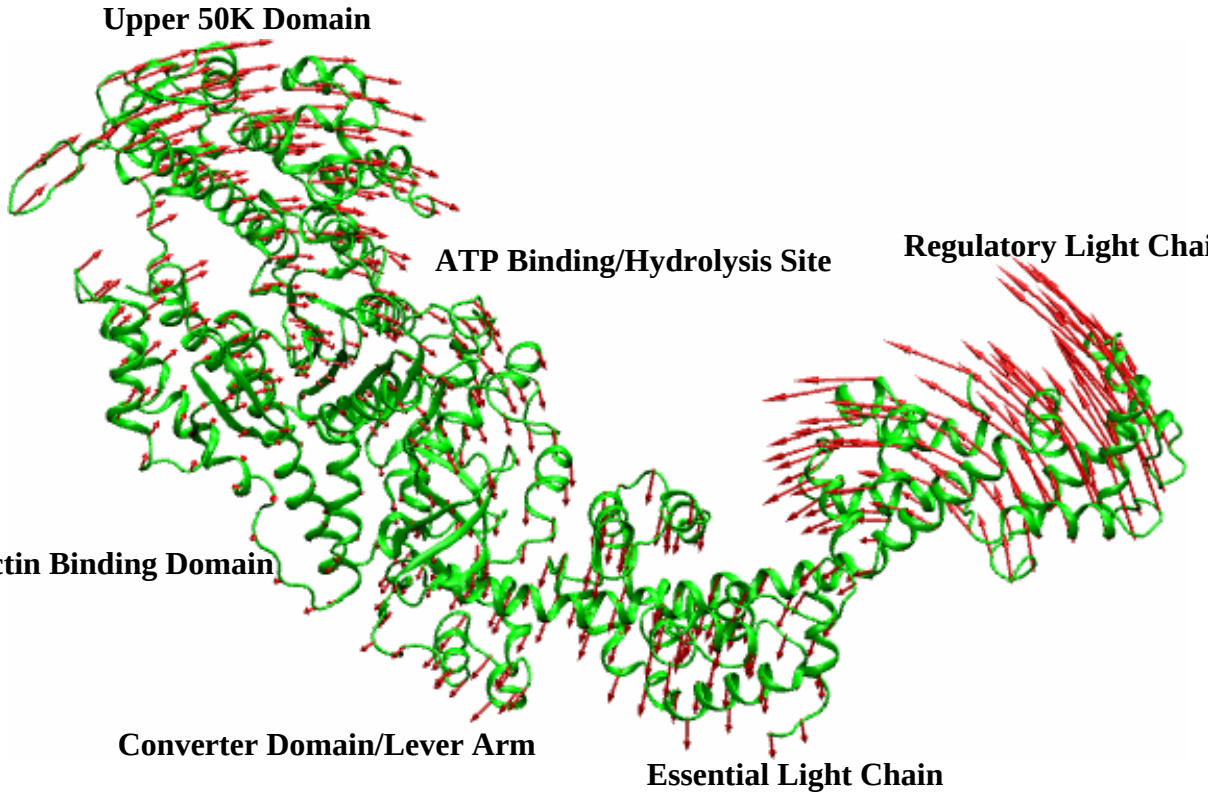


ACM: 3-ns
3 modes @ 800K
others @ 300K
Modes are updated every 100 time steps
Standard MD 3-ns all @ 300K
SPC water model

Domain motions in T4L



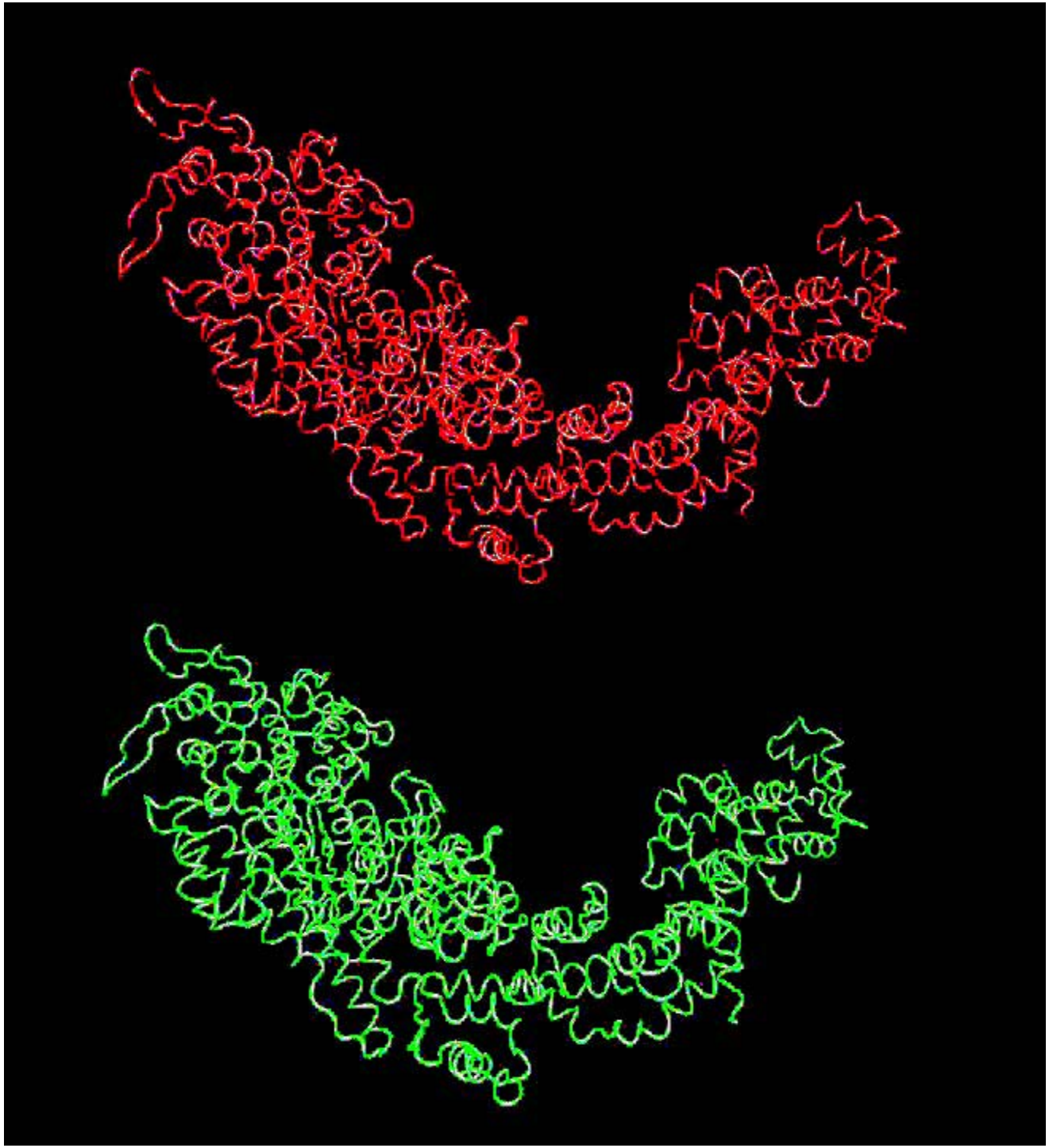
Myosin – Normal Mode Analysis



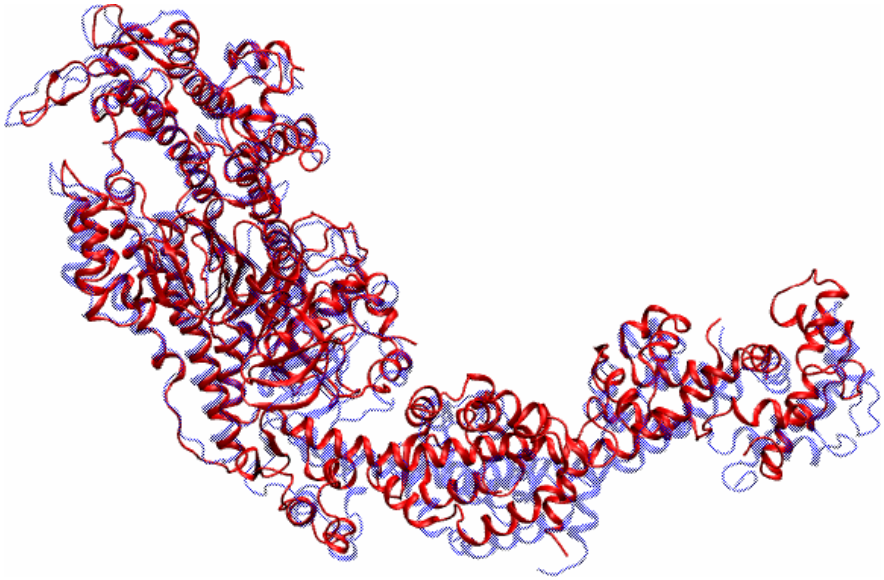
The myosin cross-bridge is a molecular machine with communicating functional units. How can the small changes at the active site be amplified into the large conformational changes? How do mutations interfere its functional dynamics?

ACM vs. MD:

Myosin

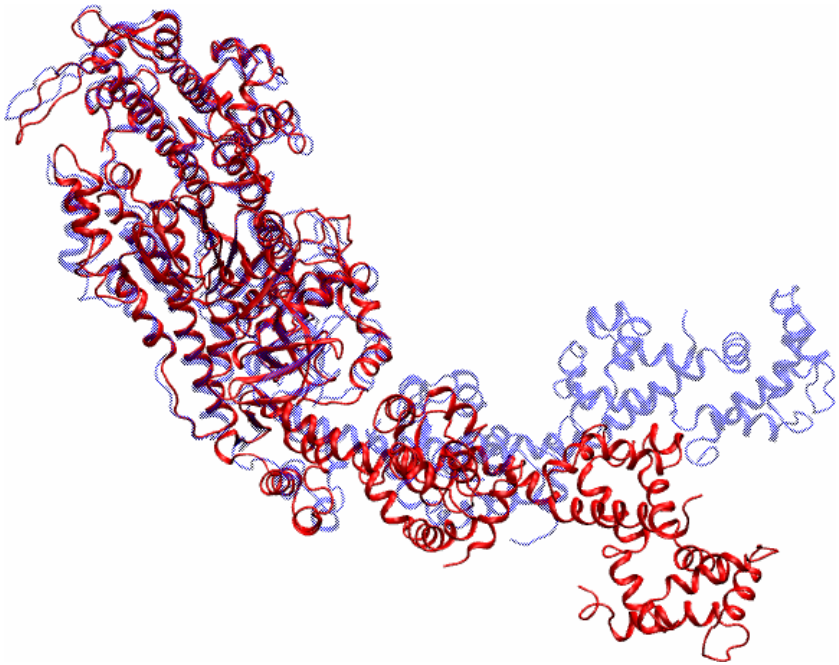


Myosin - structure comparison



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 *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):

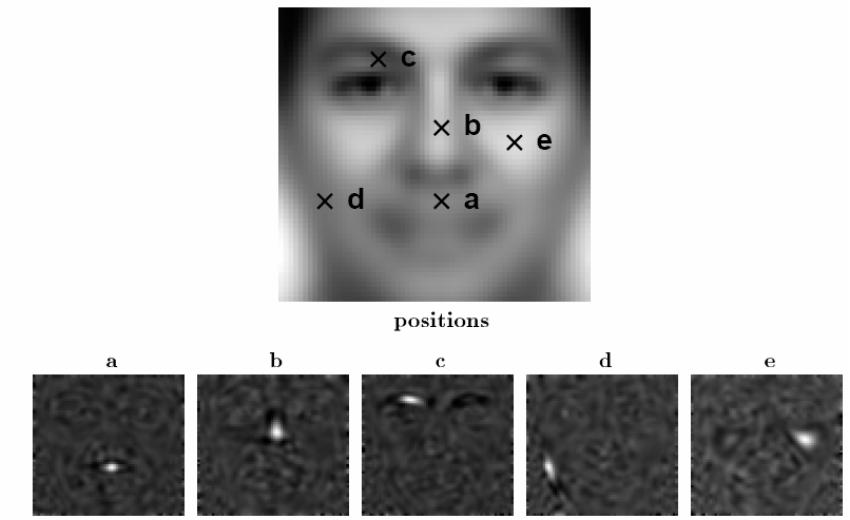
Balsera MA, Wriggers W, Oono Y, Schulten K: **Principal Component Analysis and Long Time Protein Dynamics.** *J Phys Chem* 1996, **100:** 2567-2572.

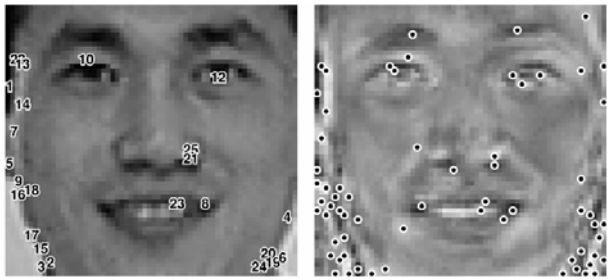
Goal: an alternative statistical theory that describe dynamic features locally and that does not suffer from the sampling and orthogonalization problems.

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?

From: Penev PS, Atick JJ: **Local Feature Analysis: A General Statistical Theory for Object Representation.** Network: computation in neural systems 1996, **7:**477-500.

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

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

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

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:

$$O^{rec}(i) = \sum_{m=1}^{|\mathcal{M}|} a_m(i) O(i_m)$$

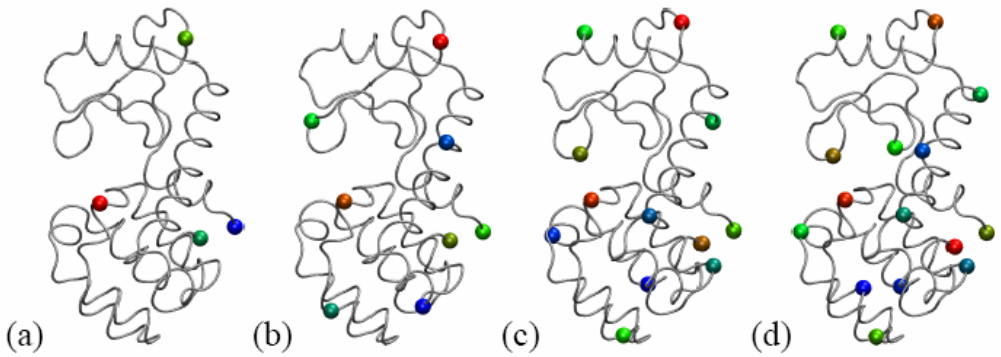
Optimal linear prediction coefficients:

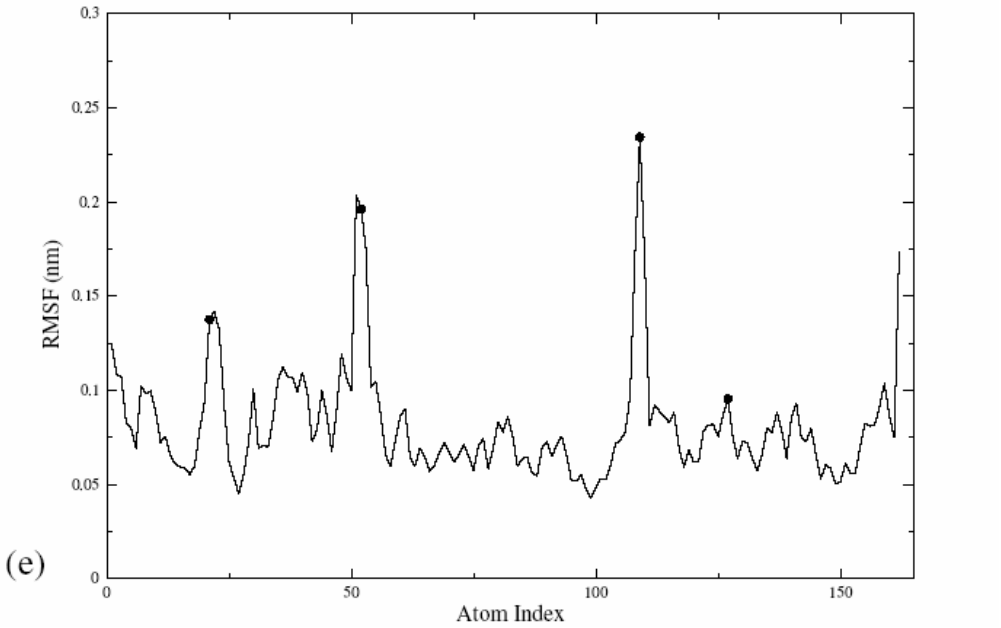
$$a_m(i) = \sum_{l=1}^{|\mathcal{M}|} P(i,\, i_l)(P'^{-1})_{lm}$$

Average reconstruction mean square error:

$$E^{rec} = \left\langle \|O^{err}(i)\|^2 \right\rangle \equiv \left\langle \|O(i) - O^{rec}(i)\|^2 \right\rangle$$

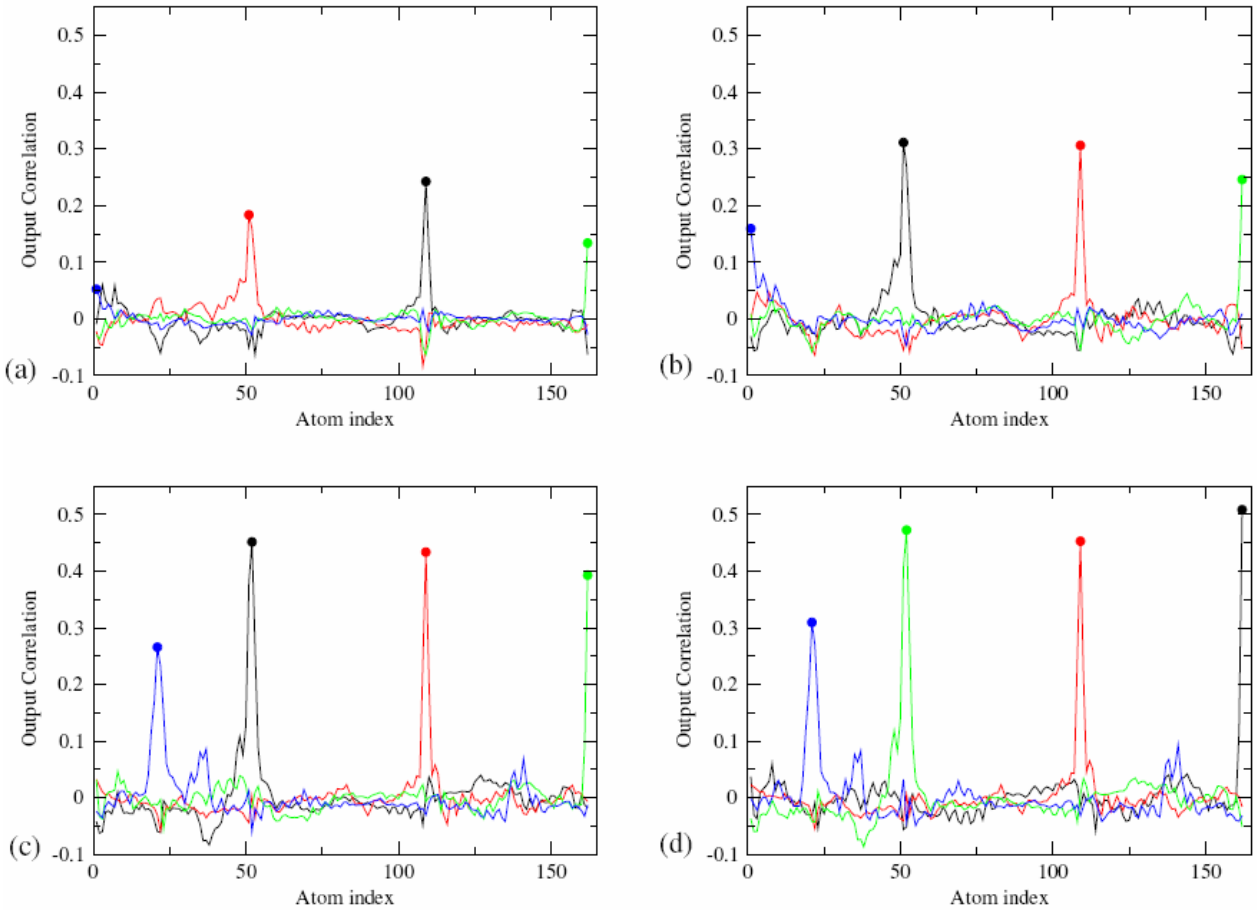
Sparse Distributions in T4L





(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.

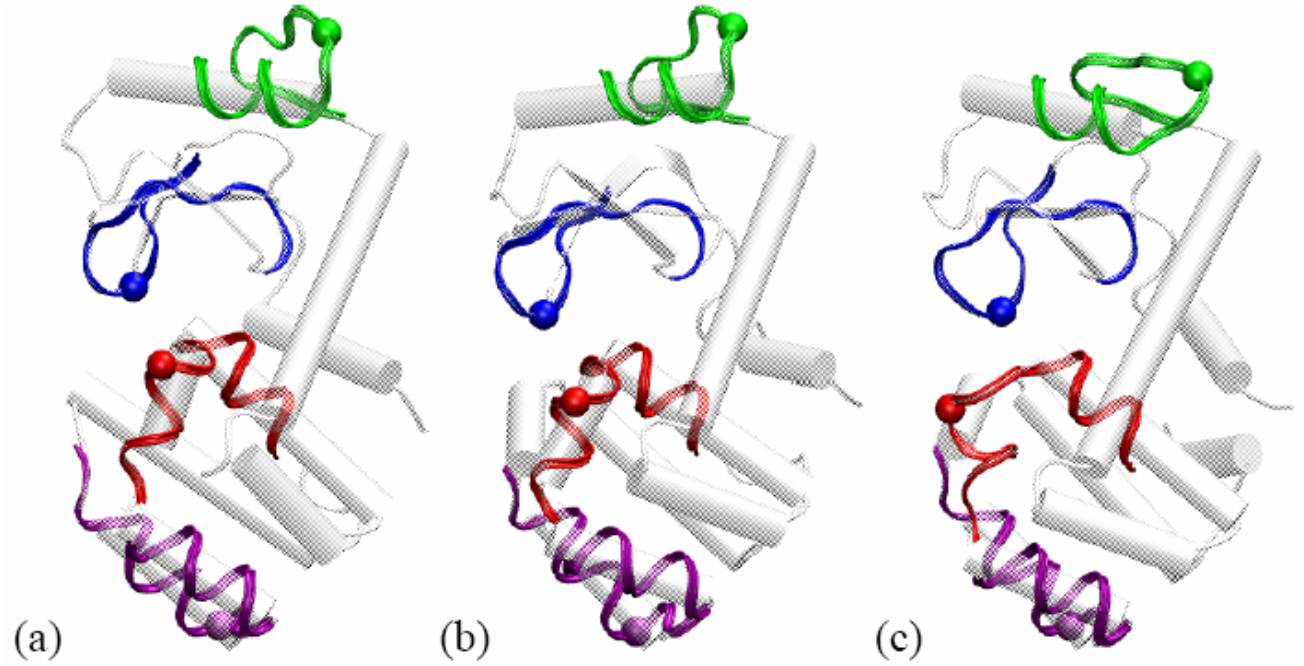
Output Functions' Correlations



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

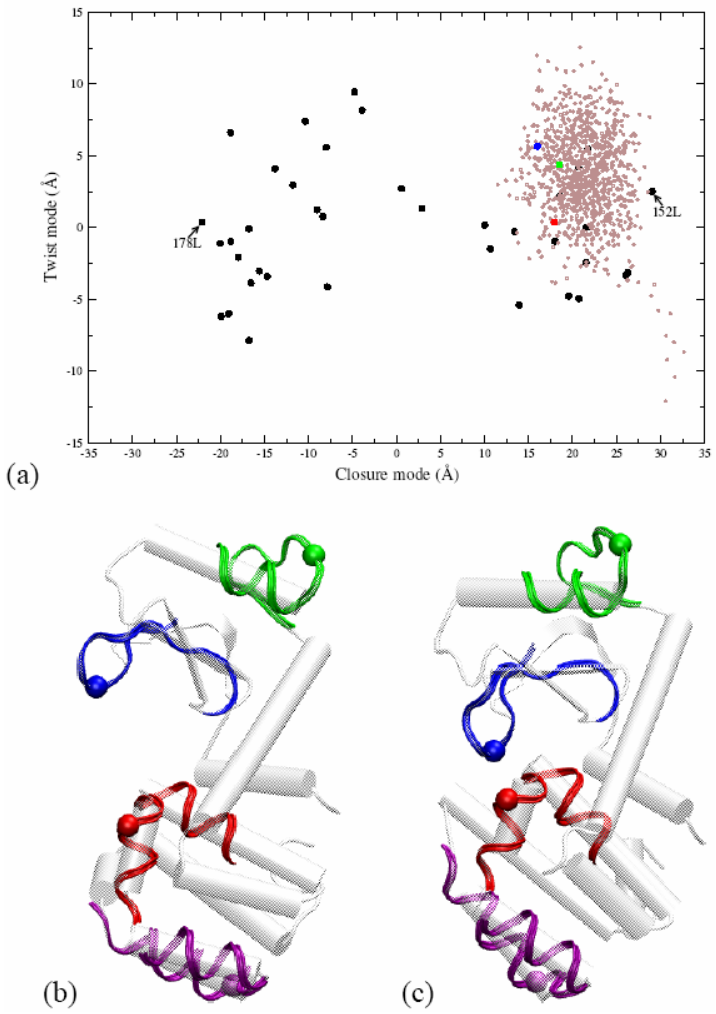
$$\left\langle \vec{O}_h \cdot \vec{O}_k \right\rangle = \sum_{d=1}^{3} \langle O(h_d) O(k_d) \rangle \equiv \sum_{d=1}^{3} P(h_d,\, k_d)$$

Local Dynamic Domains in T4L



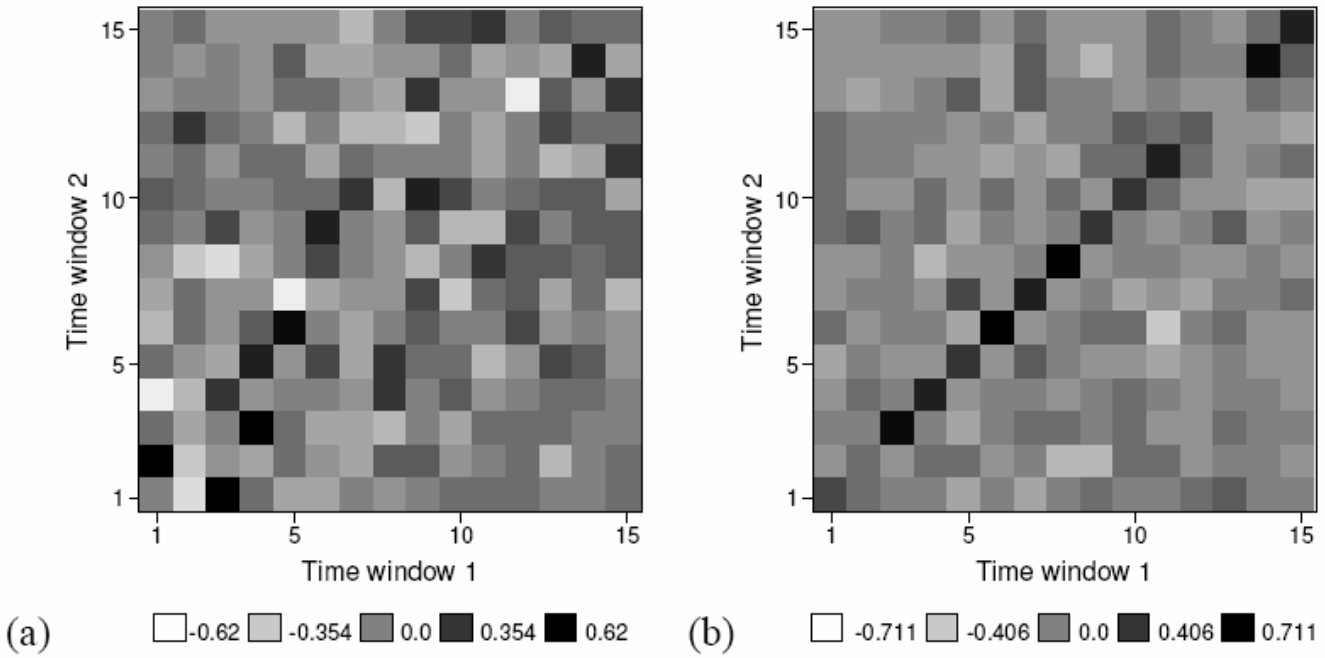
(a) t=0 ns, (b) t=4.00 ns, and (c) t=8.25 ns.
Four local features with different colors

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).

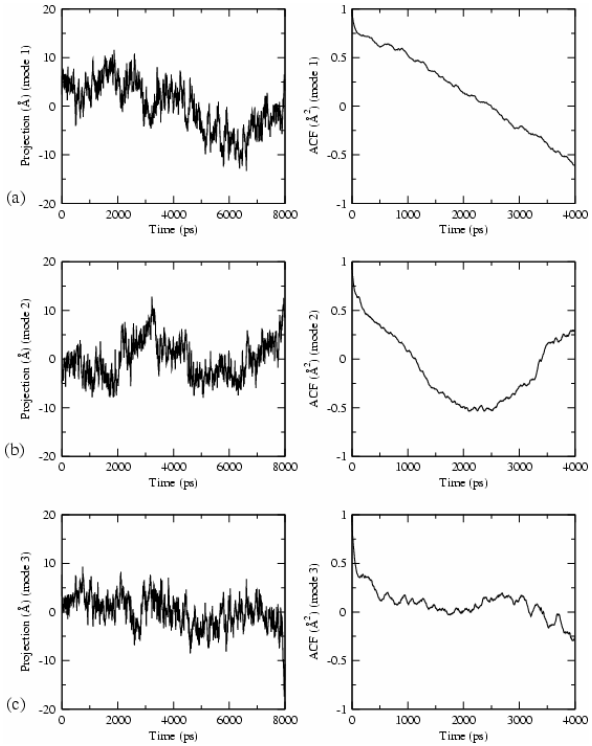
Convergence of PCA and LFA



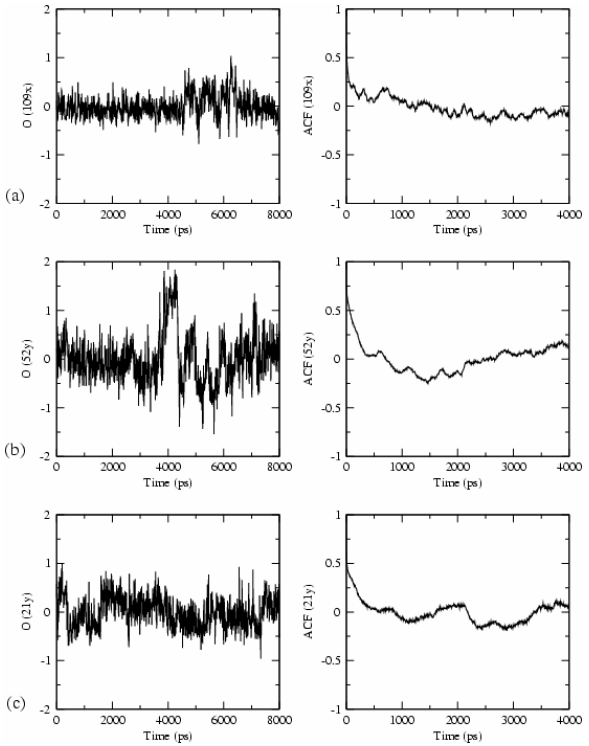
Different time windows have almost the same local features.

I[a]	II[b]	overlap[c]
162	162	0.322
21	23	0.151
1	1	0.571
52	52	0.473
30	32	0.436
127	127	0.711
69	69	0.479
40	40	0.705
136	137	0.456
116	119	0.383
92	93	0.546
107	109	0.219
10	60	-0.084
80	81	0.584
151	154	0.463

Convergence of PCA and LFA



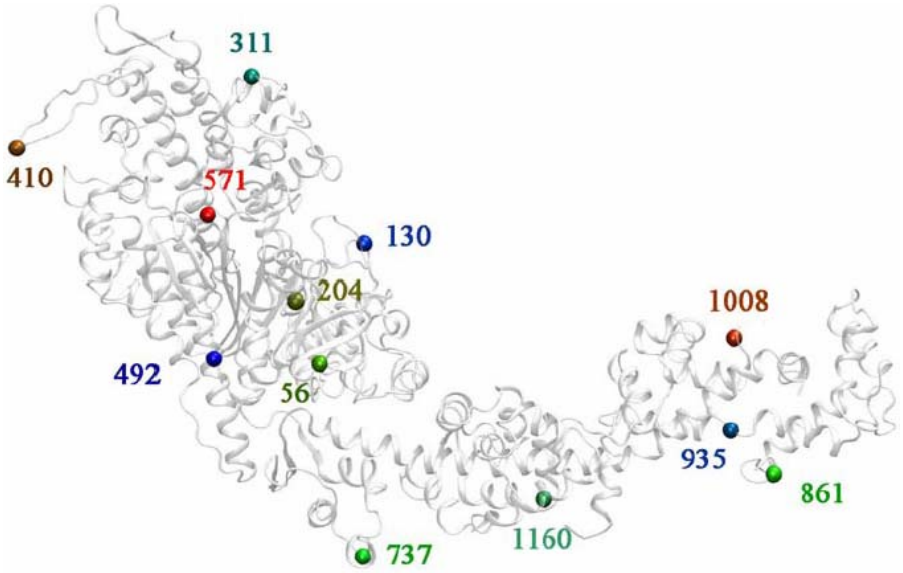
PCA output functions



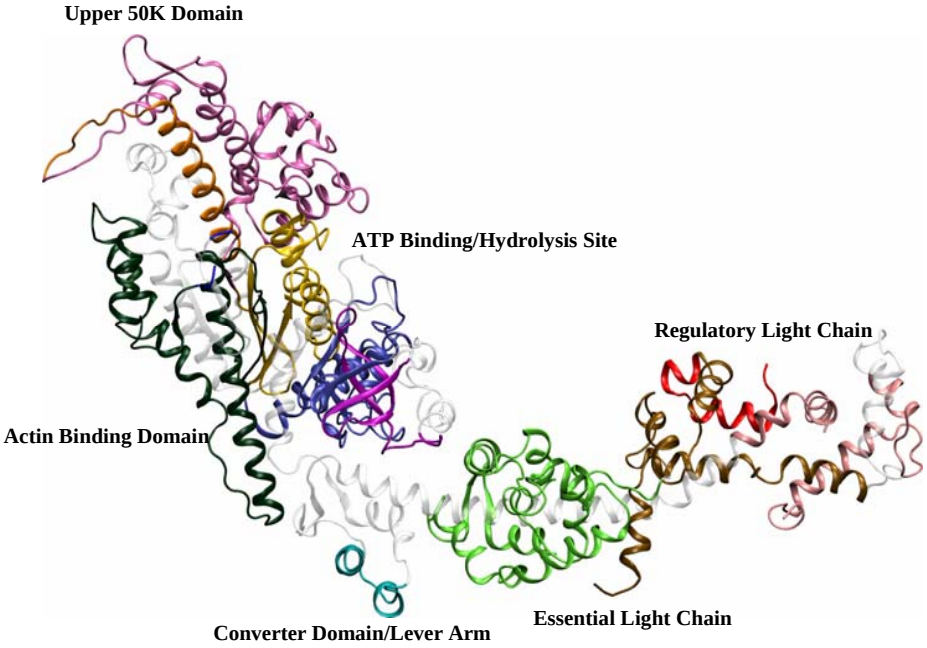
LFA output functions

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
http://kybele.psych.cornell.edu/%7Eedelman/Psych-465-Spring-2003/PCA-tutorial.pdf
Monique M Tirion (1996) Phys Rev Lett. 77:1905-1908
Zhang et al., Biophys J. (2003) 84:3583-93.
Amadei, Linsen, Berendsen, Proteins (1993), 17:412-425
Balsera, Wriggers, Oono, Schulten J. Phys. Chem. 100:2567-2572 (1996)

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).