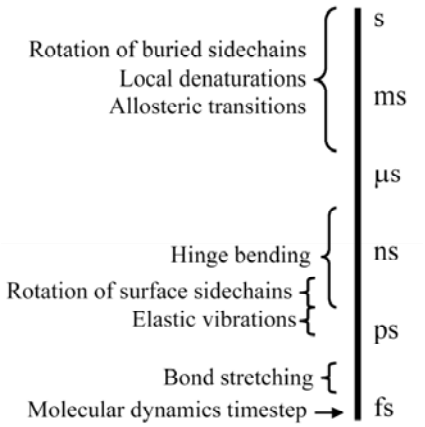
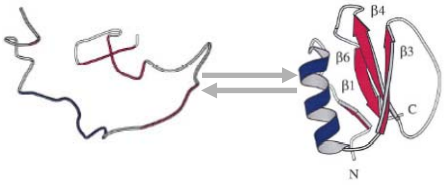
PCA and LFA

Zhiyong Zhang, Ph.D.
Willy Wriggers, Ph.D.

The Molecular Dynamics Sampling Problem





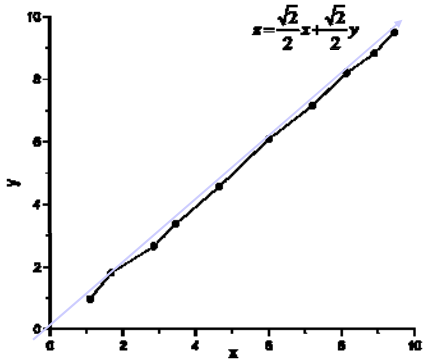
protein folding/unfolding

large-scale functional motion

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

vibration of bonds: 10^{-15} s

Collective Coordinates and Dimensionality Reduction



Collective Coordinates

- **Diagonalize Hessian matrix**

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

- **Principal Component Analysis**

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

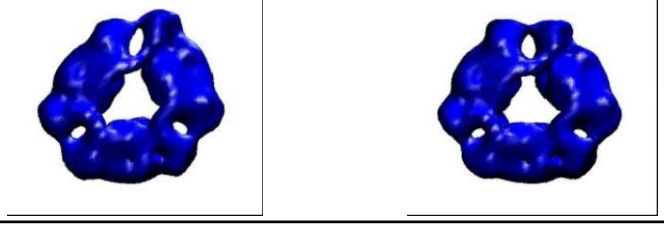
- **Normal Mode Analysis**

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

Functional motions of a protein may be represented by only a few low-frequency modes.

Global Collective Coordinates: What are the Limitations?

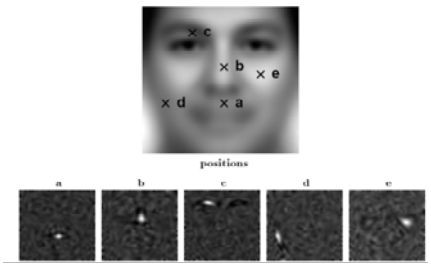
- In NMA, we do not know *a priori* which is a functionally relevant mode, the first 12 low-frequency modes are probable candidates.
- In PCA, the global modes don't converge due to time limitations of the molecular dynamics simulation (sampling problem). Balsera MA, Wriggers W, Oono Y, Schulten K: *J Phys Chem* 1996, 100: 2567-2572.
- Both methods break the symmetry of structures due to forced orthogonalization:

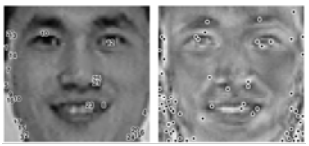


Local Feature Analysis (LFA)

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

Unlike the global eigenmodes, LFA describes objects in terms of statistically derived local features and their positions.





Is LFA applicable to protein 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: $C(i,j) \equiv \langle \Delta x_i \Delta x_j \rangle \equiv \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$

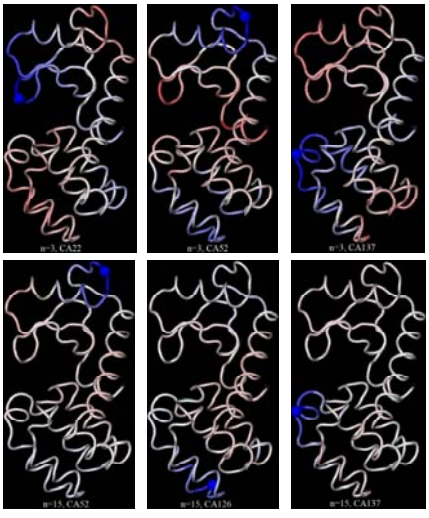
PCA: $C(i,j) = \sum_{r=1}^{3N} \Psi_r(i) \lambda_r \Psi_r(j)$ ⟶ **PCA output:** $A_r = \sum_{i=1}^{3N} \Psi_r(i) \Delta x_i \equiv \sum_{i=1}^{3N} K_r(i) \Delta x_i$

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

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: $\langle O(i) O(j) \rangle = \sum_{r=1}^{n} \Psi_r(i) \Psi_r(j) \equiv P(i,j)$

Output Correlation



n=3

n=15

Local Feature Analysis (LFA)

- Theory (II)

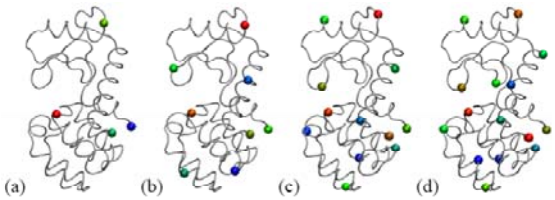
We replaced the n global PCA modes with the full 3N LFA output functions. Therefore an additional dimensionality reduction step is required in the LFA output space. We approximate the entire 3N outputs with only a small subset of them that correspond to the strongest local features by taking advantage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

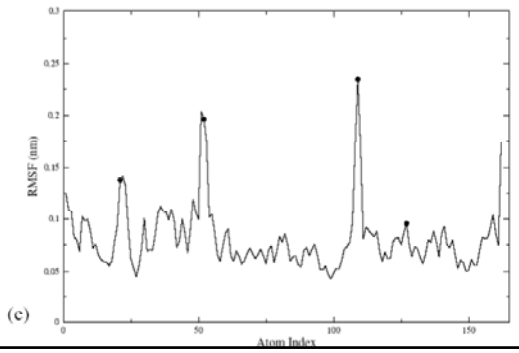
Reconstruct the outputs: $$O^{rec}(i) = \sum_{m=1}^{|\mathcal{M}|} a_m(i) O(i_m)$$

Optimal linear prediction coefficients: $$a_m(i) = \sum_{l=1}^{|\mathcal{M}|} P(i, i_l)(P^{t-1})_{lm}$$

Average reconstruction mean square error: $$E^{rec} = \left\langle \|O^{err}(i)\|^2 \right\rangle \equiv \left\langle \|O(i) - O^{rec}(i)\|^2 \right\rangle$$

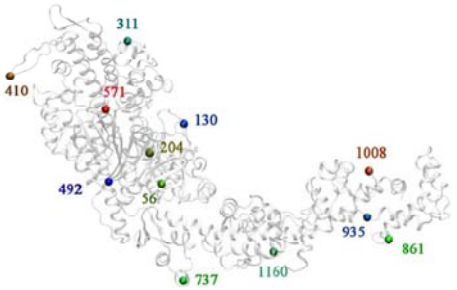
Sparsification

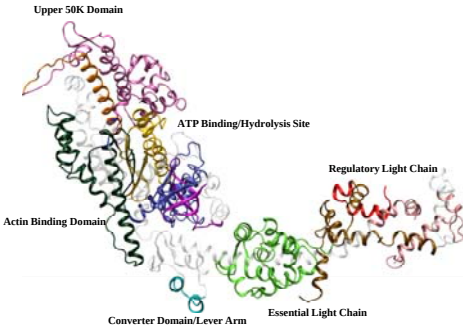




(a) The first 4 PCA modes were used to do LFA, n=4; (b) n=8, (c) n=12, and (d) n=15. (e) Root-mean-square fluctuations of C_alpha atoms in T4L.

Local Feature Analysis of Myosin



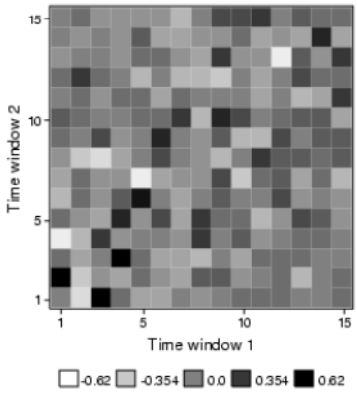


Twelve seed atoms

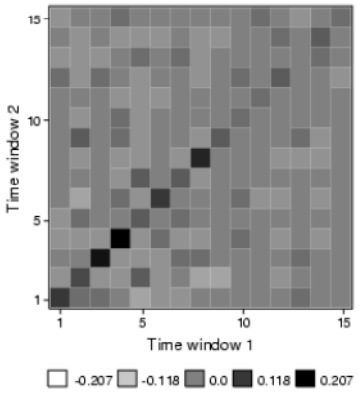
Twelve local dynamic domains

Convergence Properties

Overlap between 15 modes from first and second half of 10ns trajectory (T4 lysozyme, standard MD)

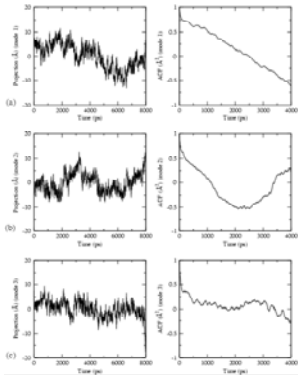


PCA

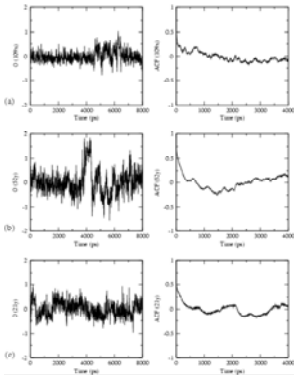


LFA

Convergence Properties



PCA autocorrelation



LFA autocorrelation

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

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.
- We perform a statistical analysis that emphasizes dynamic domains that are moving independently from each other.
- Our References:
  - *Zhiyong Zhang and Willy Wriggers*. Local Feature Analysis: A Statistical Theory for Reproducible Essential Dynamics of Large Macromolecules. *Proteins: Structure, Function, and Bioinformatics*. 2006, Vol. 64, pp. 391-403
  - *Willy Wriggers, Zhiyong Zhang, Mili Shah, and Danny C. Sorensen*. Simulating Nanoscale Functional Motions of Biomolecules. *Molecular Simulation* 2006, Vol. 32, pp. 803-815.
  - *Zhiyong Zhang and Willy Wriggers*. Coarse-Graining Protein Structures With Local Multivariate Features from Molecular Dynamics. *J. Phys. Chem. B* 2008, Vol. 112, pp. 14026-14035.