PROTEINS: Structure, Function, and Bioinformatics 64:391–403 (2006)

Local Feature Analysis: A Statistical Theory for Reproducible Essential Dynamics of Large Macromolecules

Zhiyong Zhang, and Willy Wriggers*

Laboratories for Biocomputing & Imaging, School of Health Information Sciences & Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, Texas

ABSTRACT **Multivariate statistical methods are widely used to extract functional collective motions from macromolecular molecular dynamics (MD) simulations. In principal component analysis (PCA), a covariance matrix of positional fluctuations is diagonalized to obtain orthogonal eigenvectors and corresponding eigenvalues. The first few eigenvectors usually correspond to collective modes that approximate the functional motions in the protein. However, PCA representations are globally coherent by definition and, for a large biomolecular system, do not converge on the time scales accessible to MD. Also, the forced orthogonalization of modes leads to complex dependencies that are not necessarily consistent with the symmetry of biological macromolecules and assemblies. Here, we describe for the first time the application of local feature analysis (LFA) to construct a topographic representation of functional dynamics in terms of local features. The LFA representations are low dimensional, and like PCA provide a reduced basis set for collective motions, but they are sparsely distributed and spatially localized. This yields a more reliable assignment of essential dynamics modes across different MD time windows. Also, the intrinsic dynamics of local domains is more extensively sampled than that of globally coherent PCA modes. Proteins 2006;64:391–403.**

Key words: principal component analysis; local feature analysis; molecular dynamics; bacteriophage T4 lysozyme; functional domain motions

INTRODUCTION

Molecular dynamics (MD) is an important tool in the study of the functional dynamics of proteins and macromolecular complexes.[1,2] One of the major limitations of MD is the shortness of achievable simulation times, typically of the order of tens to hundreds of nanoseconds. These times are much shorter than the time scales of many important biological processes, such as multidomain motions and allosteric transitions, that take place on the millisecond scale and beyond.[3,4] Therefore, attempts have been made to extract 'essential' functional features from the short trajectories, with the hope to describe the motion in terms of a small number of variables, sometimes called collective coordinates or essential degrees of freedom.[5–11]

One widely used statistical approach to such dimensionality reduction is principal component analysis (PCA),[12,13] also known as the Karhunen–Loeve expansion[14] in time series analysis. This statistical method was introduced to the protein research community by McCammon, Karplus, and their coworkers[15,16] in the 1980s under the name *quasi-harmonic analysis*. Since the early 1990s, PCA-based essential dynamics techniques have enjoyed the increasing enthusiasm of a large number of investigators[7,8] who successfully applied them to investigate the physical nature of protein dynamics and to sample the conformational space.[10,11]

While there is general agreement about the heuristic appeal of PCA for the prediction of functionally relevant modes, it became necessary in the mid-1990s to investigate the limitations of PCA conferred by the MD sampling problem. García and colleagues demonstrated that for large systems the distribution of conformations becomes multimodal[9] (as suggested also by Go's jumping-among-minima model[17]), leading to a breakdown of the quasi-harmonic assumption. Also, Clarage and colleagues showed that correlations in low-frequency displacements are under sampled by nanosecond MD simulations and asked the question "How long is long enough?"[18] An answer may be found in experimental studies that suggest that the relaxation times of correlations for multidomain proteins are on the order of milliseconds or longer.[3,4] This led Balsera and coworkers to conclude that PCA modes from short MD trajectories are intrinsically unreliable.[19]

Here, we take the conciliatory view that PCA may serve as a useful filter for identifying a reduced dimensional, or essential subspace, although it is clear from the prior work that individual PCA modes may overestimate the coherence of long-distance motions due to limited sampling and due to the global extent of the modes. The PCA filtering enables a subsequent local representation of the dynamics

Grant sponsor: NIH; Grant numbers: 1R01GM62968, 1R90DK071505-01; Grant sponsor: Human Frontier Science Program; Grant number: RGP0026/2003; Grant sponsor: Alfred P. Sloan Foundation; Grant number: BR-4297.

*Correspondence to: Willy Wriggers, Laboratories for Biocomputing & Imaging, School of Health Information Sciences & Institute of Molecular Medicine, University of Texas Health Science Center at Houston, 7000 Fannin St., Suite 600, Houston, TX 77030. E-mail: wriggers@biomachina.org

Received 13 October 2005; 18 January 2006; 7 February 2006

Published online 12 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20983

described below. In this filtering role of PCA, it is not necessary to know a priori which particular PCA mode (or which linear combination of modes) is functionally relevant. Our minimal assumption is that only the combined subspace is relevant, as suggested by the findings of Amadei and colleagues[7] and by a recent survey of harmonic (or normal mode) analysis of protein dynamics, where observed conformational changes are most often contained within the subspace of the first 12 low-frequency modes.[20]

The global extent of individual PCA modes is problematic not only because of the limited sampling of long-range correlations, but also because of the forced orthogonalization of the modes. Since the nth mode is always forced to be orthogonal to the first $n - 1$ modes, complex causal dependencies arise. For example, one can not claim that a particular mode n is functionally isolated from (or less relevant than) slower modes. In Balsera's paper,[19] it was shown that even fast modes, whose relaxation time is well within the MD sampling window, cannot be recovered by PCA due to their dependence on the slower, undersampled modes. The forced orthogonalization also has the undesirable effect of breaking the symmetry of large-scale macromolecular assemblies. For example, a three-fold symmetric system should exhibit a symmetry related representation (for each 120° rotation), instead PCA fixes by numeric chance one of the three possible solutions and forces all subsequent modes to be orthogonal, thereby breaking the symmetry.

Due to the apparent limitations of global collective coordinates, we were seeking an alternative statistical theory that describes dynamic features locally and that does not suffer from the sampling and orthogonalization problems. A particularly promising recent approach is non-negative matrix factorization (NMF)[21,22] of image data, which has been used for classification tasks in face recognition. Compared to the global PCA representation (eigenfaces), the NMF basis corresponds to recognizable localized features such as parts of a face (eyes, nose, ears, and mouth). Instead of the forced orthogonalization as in PCA, NMF uses non-negativity constraints in the matrix factorization, which lead to a parts-based representation of the objects. Unfortunately, this promising concept is not applicable to protein dynamics, as the elements in the covariance matrix could have either sign and can not be restricted to positive values as in gray value images.

Earlier, Penev and Atick developed an alternative statistical technique, termed *local feature analysis* (LFA), to construct a local topographic representation of objects from the global PCA modes.[23] It turns out that LFA is free from non-negativity constraints, although this was not exploited at the time. As in the case of NMF the LFA basis functions are sparsely distributed and give a description of objects in terms of local features and their positions. In this article, we adapted for the first time the theoretical framework of LFA to the study of protein dynamics. We obtained local features that clearly correspond to segmented dynamic domains in the protein. Also, LFA provides for a significant improvement in the reproducibility and convergence of the statistical sampling.

The organization of this paper is as follows. Firstly we will describe our adaptation of the theory of LFA, as well as computational details for the MD simulation of a test system, bacteriophage T4 lysozyme (T4L). Subsequently, we provide results and a discussion regarding the performance features of LFA. Finally, we provide concluding remarks on the parameterization and future applicability of the algorithm.

THEORY AND METHODS

Local Representations from PCA Modes

Assuming a protein structure, for simplicity we only consider here the coordinates of a number N Cα atoms. Amadei and colleagues demonstrated that the identity of the larger amplitude modes is robust under such Cα coarse graining.[7,24] A generalization to all atoms is straightforward. After eliminating the overall translational and rotational motion from the MD simulation as is customary in PCA, the internal motion is described by a trajectory $x(t)$, where x is a $3N$-dimensional column vector of the Cα atomic coordinates: $\{x_1, x_2, \cdots, x_{3N}\}$. The correlations of atomic fluctuations are expressed in a covariance matrix

$$C(i,j) \equiv \langle \Delta x_i \Delta x_j \rangle \equiv \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle, \quad (1)$$

where $\langle\rangle$ denotes an average over the time frames. In PCA, we diagonalize the covariance matrix to produce the orthogonal set of eigenvectors (PCA modes) $\psi_r(i)$, $r = 1, \cdots, 3N$ and corresponding eigenvalues λ_r:

$$C(i,j) = \sum_{r=1}^{3N} \psi_r(i) \lambda_r \psi_r(j). \quad (2)$$

The displacements Δx_i can then be reconstructed from the PCA modes

$$\Delta x_i = \sum_{r=1}^{3N} A_r \psi_r(i) \text{ with } A_r = \sum_{i=1}^{3N} \psi_r(i) \Delta x_i \equiv \sum_{i=1}^{3N} K_r(i) \Delta x_i, \quad (3)$$

where A_r is the so-called *output* of the representation, that is, the projection of atomic fluctuations onto the PCA mode ψ_r. PCA outputs are decorrelated in the sense that $\langle A_r A_q \rangle = \lambda_r \delta_{rq}$. $K_r(i)$ is the so-called *kernel* of the PCA representation, in the case of PCA $K_r(i) = \psi_r(i)$. We choose to sort λ_r in a decreasing order, thus the first eigenvector represent the motion that has the largest positional deviation. As explained above, we assume that a small number $n (n \ll 3N)$ of modes are sufficient to describe the dominant dynamics. This means we truncate the expansion (Eq. 3) early and define the (approximate) reconstructed deviations:

$$\Delta x_i^{rec} = \sum_{r=1}^{n} A_r \psi_r(i). \quad (4)$$

The PCA representation offers a reduced dimensionality, however, it is nonlocal. By this we mean that the kernel

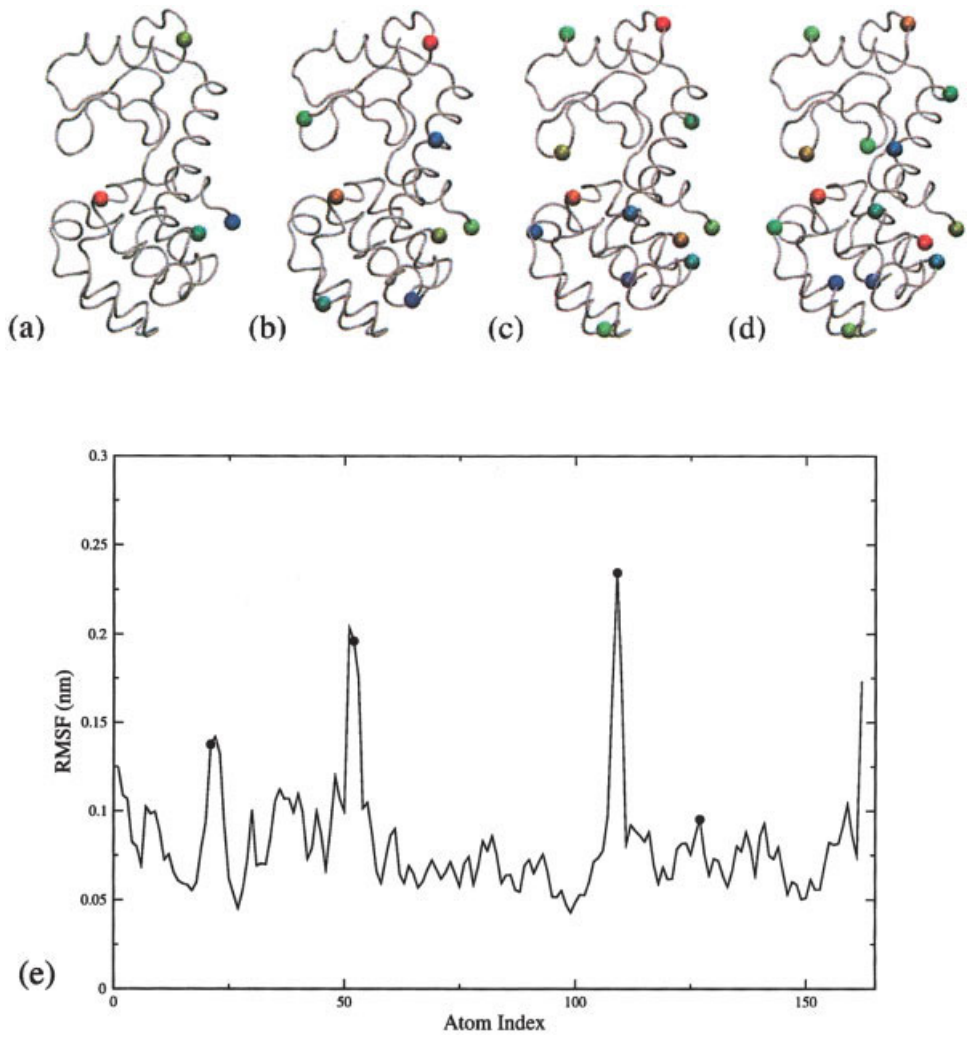

Fig. 1. Sparsification results of T4L. The first n PCA modes were used to perform LFA: (**a**) $n = 4$; (**b**) $n = 8$; (**c**) $n = 12$; (**d**) $n = 15$. The selected C_α atoms are represented by spheres. Red–green–blue colors indicate the order of selection by the algorithm (the first selected C_α atom is red, and the last one is blue). (**e**) Root-mean-square fluctuations (RMSF) of C_α atoms in T4L. Trajectory frames from 2 to 10 ns were used to calculate the RMSF. Four seed atoms ($C_\alpha - 21$, $C_\alpha - 52$, $C_\alpha - 109$, and $C_\alpha - 127$) are indicated by black dots.

functions $K_r(i)$ (Eq. 3) extend over the entire range of i (the $3N$ degrees of freedom of the protein), but nearby values in the r index have no relationship among each other. For the desired LFA, we recast the expansion into a new representation that obeys locality, that is, the kernel functions are not labeled by the PCA mode index r, but by the index of the degrees of freedom (DOF), i. The most general form for the LFA kernel is

$$K(i,j) = \sum_{r,s=1}^{n} \psi_r(i) Q_{rs} \psi_s(j), \tag{5}$$

where Q_{rs} is an arbitrary matrix. Similar to the PCA outputs A_r in Equation 3, we define local outputs $O(i)$

$$O(i) \equiv \sum_{j=1}^{3N} K(i,j)\Delta x_j = \sum_{r,s=1}^{n} \psi_r(i) Q_{rs} A_s, \tag{6}$$

but here O depends on i and not on r. We know that the PCA outputs A_r are decorrelated by frequency, so likewise we seek to decorrelate the local outputs $O(i)$ by space. Because the $3N$ outputs $O(i)$ are derived from only $n \ll 3N$ linearly independent A_r, the decorrelation condition $\langle O(i)O(j)\rangle = \delta(i,j)$ is no longer satisfied. Instead we seek to satisfy the condition of minimum correlation of the outputs $O(i)$ by minimizing the mean-square deviation

$$msd = \sum_{i,j=1}^{3N} |\langle O(i)O(j)\rangle - \delta(i,j)|^2 \tag{7}$$

with respect to the matrix Q. One can show that Q must be given by $Q_{rs} = \frac{1}{\sqrt{\lambda_r}} U_{rs}$,[23] and U_{rs} is any orthogonal matrix satisfying $U^{\mathrm{T}}U = 1$. In general, a variety of U_{rs} may be employed while preserving the decorrelation,[25] but here

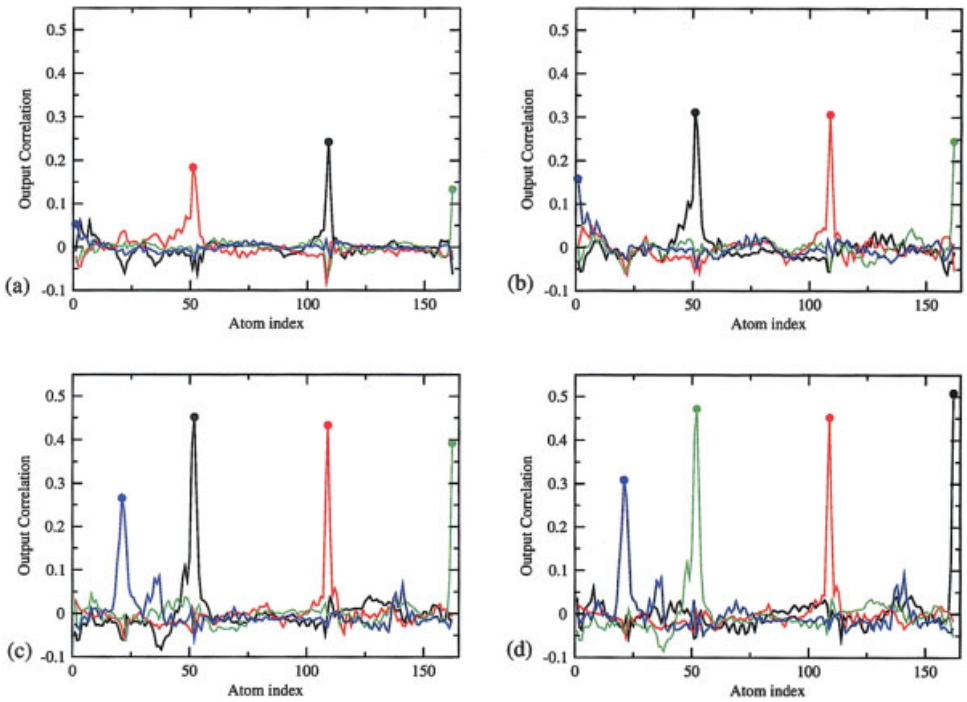

Fig. 2. Output correlations (Eq. 16) between the seed atoms and the protein as a function of residue number. Colors indicate selection orders of seed atoms: (1) black, (2) red, (3) green, and (4) blue. (**a**) $n = 4$. Black, $C_\alpha - 109$; red, $C_\alpha - 51$; green, $C_\alpha - 162$; and blue, $C_\alpha - 1$. (**b**) $n = 8$. Black, $C_\alpha - 51$; red, $C_\alpha - 109$; green, $C_\alpha - 162$; and blue, $C_\alpha - 1$. (**c**) $n = 12$. Black, $C_\alpha - 52$; red, $C_\alpha - 109$; green, $C_\alpha - 162$; and blue, $C_\alpha - 21$. (**d**) $n = 15$. Black, $C_\alpha - 162$; red, $C_\alpha - 109$; green, $C_\alpha - 52$; and blue, $C_\alpha - 21$.

we consider only the simplest choice $U_{rs} = \delta_{rs}$ to remain consistent with the PCA subspace. Thus, according to Equation 6 the LFA outputs become

$$O(i) = \sum_{j=1}^{3N} \left(\sum_{r=1}^{n} \psi_r(i) \frac{1}{\sqrt{\lambda_r}} \psi_r(j) \right) \Delta x_j = \sum_{r=1}^{n} \frac{A_r}{\sqrt{\lambda_r}} \psi_r(i) \tag{8}$$

and their residual correlation is given by

$$\langle O(i)O(j) \rangle = \sum_{r=1}^{n} \psi_r(i)\psi_r(j) \equiv P(i,j). \tag{9}$$

In the limit $n \to 3N$ the LFA outputs are completely decorrelated: $P(i, j) \to \delta(i, j)$. Finally, we have derived a new kernel function

$$K(i,j) = \sum_{r=1}^{n} \psi_r(i) \frac{1}{\sqrt{\lambda_r}} \psi_r(j), \tag{10}$$

which satisfies locality. We reconstruct Δx_i using these kernels.

The matrices K and P are central to LFA and one can derive from the above equations some additional noteworthy features. First, K is by definition the projection operator onto a local feature. In the limit $n \to 3N$, the resulting dimensionless projections (or outputs) $O(i)$ become orthogonal, as well as normalized to unity, as square-integrable functions over the time domain. Second, it is straightforward to show from Equations 4 and 9 that

$$\sum_{j=1}^{3N} P(i,j)\Delta x_j = \Delta x_i^{rec}. \tag{11}$$

This means that P serves a dual role both as the correlation of the results of K (the LFA outputs) and as the projection operator onto the low-frequency subspace spanned by n PCA modes. Used as a projection operator the results of P then have length units (unlike the results of K that are dimensionless). Third, from Equations 4 and 8 we get

$$\Delta x_i^{rec} \equiv \sum_{j=1}^{3N} K^{(-1)}(i,j)O(j), \tag{12}$$

where $K^{(-1)} = \sum_{r=1}^{n} \psi_r(i) \sqrt{\lambda_r} \psi_r(j)$ is the so-called *reconstructor* or inverse kernel of the representation.

In summary, the above LFA theory can be formulated in a compact form as follows. If we define a family of functions $K^{(m)}(i, j) = \sum_{r=1}^{n} \psi_r(i) \left(\frac{1}{\sqrt{\lambda_r}} \right)^m \psi_r(j)$, then it follows that $K^{(1)}(i,j) \equiv K(i, j)$ is the LFA kernel (Eq. 10); $K^{(0)}(i, j) \equiv P(i, j)$ is the residual output correlation (Eq. 9); $K^{(-1)}(i, j)$ is the reconstructor (Eq. 12); and $K^{(-2)}(i, j) \equiv C(i, j)$ is the covariance matrix (Eq. 2).

Sparsification from Local Features

In the previous section, we replaced the n global PCA modes with a much larger number $3N$ of local LFA output functions $O(i)$. Although locality was achieved, it came at a price of expanding again to the full number of DOF. Therefore, an additional dimensionality reduction step is required in the LFA output space. This sparsification takes advantage of the fact that neighboring outputs are highly correlated. We approximate the entire $3N$ outputs $O(i)$ with only a small subset of $\{O(i_m)\}_{i,m\in\mathcal{M}}$ that correspond to the strongest local features. The other $O(i)$ can then be reasonably well predicted via the correlations $P(i, i_m) \equiv P_m(i)$.

We begin with an empty set $\mathcal{M}^{(0)} = \{0\}$. At each step, out of the $3N$ total DOF we add a seed index to $\mathcal{M}$, chosen according to the criteria described below, the seed index corresponds to either x, y, or z, coordinates of a given seed atom. Given the current set $\mathcal{M}^{(m)}$, we can reconstruct the outputs:

$$O^{rec}(i) = \sum_{m=1}^{|\mathcal{M}|} \mathrm{a_m(i)O(i_m)}. \quad (13)$$

One can show[23] that the optimal linear prediction coefficients $a_m(i)$, defined to minimize the average reconstruction mean square error on $O(i)$

$$E^{rec} = \langle \|O^{err}(i)\|^2 \rangle \equiv \langle \|O(i) - O^{rec}(i)\|^2 \rangle, \quad (14)$$

are given by

$$a_m(i) = \sum_{l=1}^{|\mathcal{M}|} P(i, i_l)(P'^{-1})_{lm}, \quad (15)$$

where P'^{-1} is the inverse of a submatrix P' from P: $P'_{lm} \equiv P(i_l, i_m)$. Out of the $3N$ available DOF we chose the seed index that has the maximum reconstruction error $O^{err}(i_{m+1})$ as the $(m+1)$th index into $\mathcal{M}$, under the condition that the seed atom and its nearest neighbors are distinct from atoms corresponding to previously found indices. We keep adding seed indices to $\mathcal{M}$ until n indices are chosen (the entire set of $O(i)$ is reconstructed without error at this time).

In principle, any n seed indices can recover the $O(i)$ without error. However, if we choose indices whose $P(i, j)$ overlap significantly, the $\{O(i_m)\}_{i_m\in\mathcal{M}}$ will be correlated and the representation would be redundant in some regions but insufficient in other regions. In the above sparsification algorithm, we choose an index whose output is predicted worst by the already chosen ones. This assures that the corresponding atom is dynamically decorrelated from the atoms corresponding to already chosen indices.

COMPUTATIONAL DETAILS

The MD simulation and some of the subsequent analysis were performed using the GROMACS package (version 3.1.4), using the GROMACS forcefield with united-atom model.[26,27] We selected bacteriophage T4 lysozyme (T4L) as a test system. T4L is composed of two domains connected by a long α-helix, there is a deep opening between the N-terminal and C-terminal domains, which is the active-site cleft.[28] There are many experimental structures of T4L and its mutants that indicate a hinge-bending–type domain motion.[29,30]

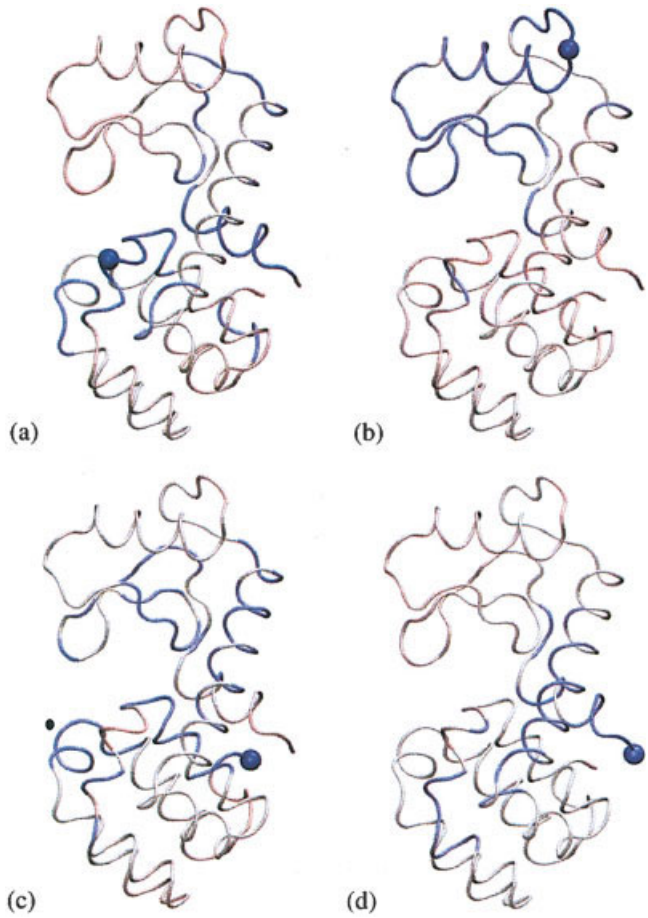

Fig. 3. T4L structures colored by output correlations (Eq. 16) between the seed atom (represented by a blue sphere) and other C_α atoms. Blue indicates positive, white indicates 0, and red indicates negative correlation values. $n = 4$ PCA modes were used for LFA, and the four seed atoms are shown by in order of selection. (**a**) $C_\alpha - 109$, (**b**) $C_\alpha - 51$, (**c**) $C_\alpha - 162$, and (**d**) $C_\alpha - 1$.

The crystal structure of T4L (PDB entry:2LZM) determined at 1.7 Å resolution was used as the starting structure.[31] Rectangular periodic boundary conditions were used with box length of 6.356 nm × 6.287 nm × 7.303 nm (the minimum distance between the solute and the box boundary is 1.2 nm). SPC water molecules were added from an equilibrated cubic box containing 216 water molecules.[32] The system, protein and water, was initially energy-minimized using the steepest descent method, until the maximum force on the atoms is smaller than 1000 kJ mol^{-1} nm^{-1}. Eight Cl^- ions were added to compensate the net positive charge on the protein, and these ions were introduced by replacing water molecules with the most favorable electrostatic potential. The energy was again minimized using the conjugate gradient algorithm, until the maximum force is below 200 kJ mol^{-1}

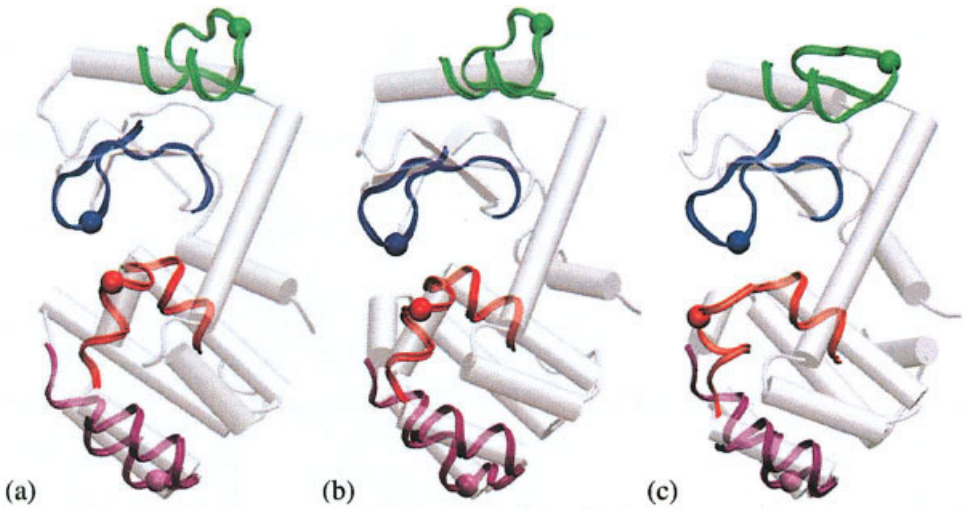

Fig. 4. Locations and dynamics of the local features in T4L during the simulation. (**a**) Initial structure of the simulation (t = 0 ns), (**b**) t = 4.00 ns, and (**c**) t = 8.25 ns. The first 15 PCA modes were used for LFA (n = 15). Four local features are colored by red (C_α − 109), green (C_α − 52), blue (C_α − 21), and purple (C_α − 127), respectively. Seed atoms are represented by spheres. The white cartoon representation indicates the secondary structure elements.

nm^{-1}. The final system contains 1683 protein atoms, 8 Cl^- ions, and 8775 water molecules, leading to a total size of 28,016 atoms. A 100 ps positional-restraint equilibration simulation was performed, with force constants 1000 kJ mol^{-1} nm^{-2}, and then followed by a 10 ns production run. The last 8 ns of this run were used to perform PCA.

We used an isothermal–isobaric simulation algorithm.[33] The three groups (protein, ions, and solvent) were coupled separately to a temperature bath of reference temperature 300 K (relaxation time 0.1 ps). The pressure was also kept constant by weak coupling to a reference value P_0 = 1 bar (relaxation time 1.0 ps). Covalent bonds in the protein were constrained using the LINCS algorithm.[34] van der Waals interactions were treated using twin-range cutoff radii (0.9 nm and 1.4 nm), and the pairlist was updated every 10 fs. The long-range electrostatic interactions were evaluated by using the particle mesh Ewald (PME) method[35] with a PME tolerance of 10^{-5} and a PME interpolation order of 4.

RESULTS AND DISCUSSION

Local Dynamic Domains in T4 Lysozyme

The first n = 4, 8, 12, and 15 PCA modes were used to construct the LFA matrices $P(i, j)$ (Eq. 9) and $K(i, j)$ (Eq. 10). This was followed by the sparsification algorithm described above to select n seed atoms. Results are visualized using VMD[36] (Fig. 1). The location and the order of the selected atoms indicate that they are allocated predominantly at the most flexible regions of the protein, in close agreements with the peaks of the root-mean-square fluctuations (RMSF) of C_α atoms [Fig. 1(e)]. The flexible N-terminal and C-terminal atoms are selected in the four cases (Fig. 1) owing to their structural variability. Other seed atoms, such as C_α − 21, C_α − 52, C_α − 109, and C_α − 127 are also selected frequently (Fig. 1) for a larger number of seed atoms selected. The functional motion of these regions is interpreted in more detail below.

We intend to represent a local feature by one seed atom and its neighboring correlated region (dynamic domain). Considering a seed atom h, its LFA output ($\vec{O}_h$) has three

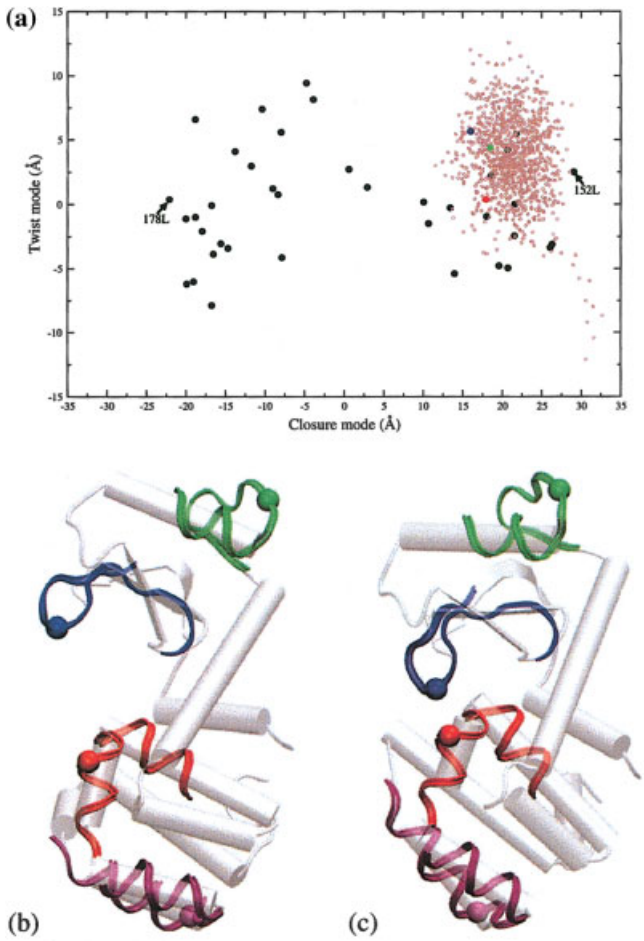

Fig. 5. (**a**) Two-dimensional projections of T4L structures from the PDB and of the MD trajectory frames onto the subspace defined by the closure and twist modes. Thirtyeight crystal structures are indicated as black dots. The extreme open (PDB entry: 178L) and closed (PDB entry: 152L) structures are labeled. Trajectory frames (see text) are shown as brown dots, except for the configurations at t = 0 ns, t = 4.00 ns, and t = 8.25 ns that are emphasized by red, green, and blue dots, respectively. (**b**) The open structure (PDB entry: 178L), and (**c**) the closed structure (PDB entry: 152L), with local features from the MD simulation highlighted as in Figure 4.

components: $O(h_d)$, $d = 1$, 2, or 3. The correlation between the seed atom and any other atom k (with LFA output $\vec{O}_k$) is:

$$\langle \vec{O}_h \cdot \vec{O}_k \rangle = \sum_{d=1}^{3} \langle O(h_d) O(k_d) \rangle \equiv \sum_{d=1}^{3} P(h_d, k_d). \quad (16)$$

Therefore, the correlation between any two atoms is represented by a 3×3 submatrix within $P(i, j)$. According to Equation 16, the correlation coefficient between LFA outputs of two atoms is the trace of this submatrix. We plot these output correlations in Figure 2 for the first four seed atoms found, and superimpose them in color onto the structure of T4L, for the special case $n = 4$, in Figure 3. It can be seen from Figures 2 and 3, the seed atoms are surrounded by prominent, spatially contiguous regions of high positive correlation. As can be expected from the asymptotic behavior of $P(i, j)$, these dynamic domains shrink in size and the peaks become sharper and higher in amplitude with increasing n (Fig. 2). The user-defined parameter n thus provides control over the size and number of the desired dynamic domains.

In the originally envisioned applications of LFA in image processing,[23] the elements of the P and K matrices are all positive. However, the elements of the covariance matrix (Eq. 1) may be positive or negative in our case. Therefore, we can observe a certain background noise level of small amplitude positive and also negative correlations outside of the dynamic domains (Figs. 2 and 3).

Defining a dynamic domain as the contiguous atoms that have positive correlations with a seed atom, we have identified the four local features associated with $C_\alpha - 21$, $C_\alpha - 52$, $C_\alpha - 109$, and $C\alpha - 127$, respectively. In Figure 4, we highlight the four dynamic regions in the start structure and two selected time frames of the simulation. Our goal was to identify similarities with the known functional dynamics observed in T4L.

Both experimental and theoretical studies reveal that T4L exhibits prominent open–close and twist motions between the two major domains.[29,30,37,38] More than 200 T4L structures have been deposited in the PDB, which provide an ensemble of accessible conformations under physiological conditions.[30] As a representation of this ensemble, a subset of 21 PDB entries with 38 unique structures was selected.[37,38] PCA on this subset of conformations indicates that the first two principal modes contribute more than 90% to the total fluctuations. The first mode (closure mode) corresponds to a open–close motion defined by an effective hinge axis perpendicular to the line connecting the centers of mass of the two domains. The second mode (twist mode) consists of a propeller twist about the line connecting the two centers of mass.

We projected the 38 experimental structures onto the two-dimensional (2D) subspace defined by the two experimentally observed modes. The structures to the left in Figure 5(a) are open conformations, whereas the structures to the right are closed. In Figure 5(b, c), we highlight the four local features in the most open (PDB entry: 178L) and the most closed configuration (PDB entry: 152L), respectively. Similar to the simulation results (Fig. 4), it can be seen that these local features participate in the experimentally observed functional domain motions in T4L.

We also projected the MD simulation trajectory frames onto the 2D subspace defined by the experimental closure and twist modes in Figure 5(a). The trajectory is confined to a small region of the experimentally accessible conformational space due to the limited sampling in the MD simulation. Nevertheless, it is remarkable that LFA from the short (10 ns) trajectory can identify the important local domains that facilitate the much larger experimentally observed variability of the structure.

The local features corresponding to $C_\alpha - 21$ and $C_\alpha - 109$ include the cross-domain active site of T4L, whereas $C_\alpha - 52$ is located at the hinge bending region between the two domains (Fig. 4). $C\alpha - 127$ corresponds to two helices that move as a single rigid body [Figs. 4, 5(b, c)]. In the following we describe a detailed statistical analysis to analyze the functional dynamics of these regions and to compare the LFA results to the standard PCA method.

LFA and PCA Mode Overlap

It was shown earlier that individual PCA modes obtained from MD simulations do not converge well within the short simulation times,[19] that is, the dominant modes change from one sampling time window to another and the modes obtained by PCA cannot predict long-time protein dynamics. To test the robustness of LFA we compare the overlap of both PCA and LFA modes across two time windows. We partitioned the trajectory into two time windows, 2 to 6 ns (**I**), and 6 to 10 ns (**II**) and performed PCA and LFA, as described above, in each window.

To compare the PCA modes we calculate the inner product

$$IP_{rs}^{PCA} = \sum_{i=1}^{3N} \psi_r^{\mathbf{I}}(i) \psi_s^{\mathbf{II}}(i) \equiv \sum_{i=1}^{3N} K_r^{\mathbf{I}}(i) K_s^{\mathbf{II}}(i), \quad (17)$$

where $\psi_r^{\mathbf{I}}$ is the rth mode obtained from window **I,** and $\psi_s^{\mathbf{II}}$ is the sth mode from window **II.** Because each mode is normalized in PCA, the inner product is unity when the two modes are identical. For the LFA modes we defined the overlap likewise as the inner product of the kernels,

$$IP_{hk}^{LFA} = \sum_{d=1}^{3} \sum_{i=1}^{3N} K^{\mathbf{I}}(h_d, i) K^{\mathbf{II}}(k_d, i)), \quad (18)$$

where $K^{\mathbf{I}}(h_d)$ is a row corresponding to atom h in matrix K from window **I,** and $K^{\mathbf{II}}(k_d)$ is a row corresponding to the atom k in matrix K from window **II.** Each atom has three rows in the matrix, so we summed them up. Since a row in K is not normalized (because of $n \ll 3N$), the actual value range of Equation 18 depends on how many PCA modes (n) are used to construct K. For the comparison with the PCA modes we renormalized these local feature basis vectors before calculating their overlaps.

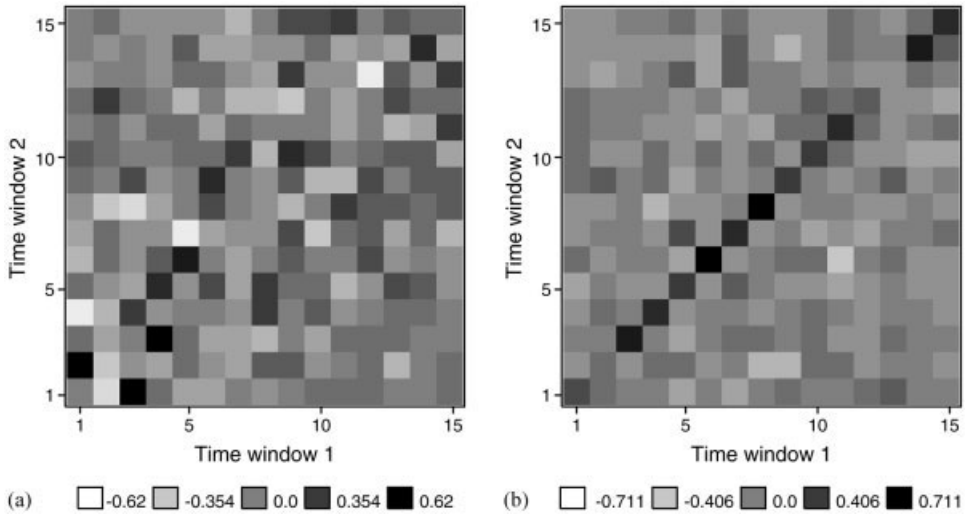

Fig. 6. (**a**) Overlaps between the PCA modes from the two time windows (Eq. 17). The first 15 PCA modes were computed and sorted by descending eigenvalues. (**b**) Overlaps between the local features (represented by seed atoms) from the two time windows (Eq. 18). Fifteen seed atoms were computed for each time window. The atoms in both time windows are sorted by the selection order of atoms in window **I.**

Figure 6(a) shows the overlap matrix between the two sets of PCA modes. If the PCA modes would converge well, the matrix should be nearly diagonal or block diagonal. However, this is not the case, as was also shown earlier,[19] due to the MD sampling problem. Although there is sometimes overlap between nearby modes [Fig. 6(a)], it is not possible to rearrange the PCA modes to make the matrix more diagonal, due to the significant differences in the associated eigenvalues. For example, the eigenvalue of mode 3 in window **I** (3 − **I**) is 7 Å^2, but that of 1 − **II** is 24 Å^2, even though both modes do overlap significantly. Likewise, the eigenvalue 1 − **I** is 24 Å^2, whereas the eigenvalue of the apparently similar mode 2 − **II** is 15 Å^2. Overall, the dominant modes are not conserved among MD time windows, and there are also new modes observed in either time window that are not observed in the other.

In Figure 6(b) we plot a similar overlap matrix for LFA, demonstrating a dramatically improved diagonal property. In the LFA matrix we can reorder the local features in window **II** according to the local features in window **I** because the LFA kernel functions $K(i,j)$ are normalized by the eigenvalues (Eq. 10). Both Figure 6(b) and Table I indicate that the two time windows exhibit very similar local features: 14 out of 15 seed atoms among the two windows are the same or very close and have significant positive overlaps. The dominant local feature in window **I** is the C-terminal C_α − 162, which has an overlap of 0.322 with window **II,** whereas the N-terminal seed atom C_α − 1 (ranked third in window **I**) exhibits a higher overlap of 0.571. This indicates that the N-terminus, although less variable [Fig. 1(e)], is more consistently sampled across time windows. Also, the rigid-body double helix corresponding to C_α − 127 exhibits an overlap of 0.711, which indicates a very good convergence of the sampling.

There are also some exceptions to the diagonal structure of the LFA basis functions (Fig. 6 and Table I). As mentioned in the above section, the local features C_α − 21 and C_α − 109 (and their corresponding seed atoms in window **II**) are near the active site of T4L related to the open–close domain motion. The dynamics of these local features is not sufficiently well sampled, because it is coupled to the large scale motion of the molecule that is under sampled in the simulation relative to the experimental variability of the structure (cf., Figs. 4 and 5). There is also one left-over feature in both cases, C_α − 10 (**I**) and C_α − 60 (**II**), which does not correspond well to that of the other time window and exhibited near-zero overlap (Table I), but this selection is less significant in terms of overall dynamics [Fig. 1(e)] and probably influenced by noise.

In summary, almost all LFA modes are well converged, in stark contrast to PCA where none of the individual modes are converged. The few outliers can be attributed to

TABLE I. Overlaps between Local Features in the Different Time Windows

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

[a]The seed atoms in window **I** are sorted by the order of selection.
[b]The seed atoms in window **II** are sorted accordingly.
[c]These overlap values are the diagonal elements in Figure 6 (b).

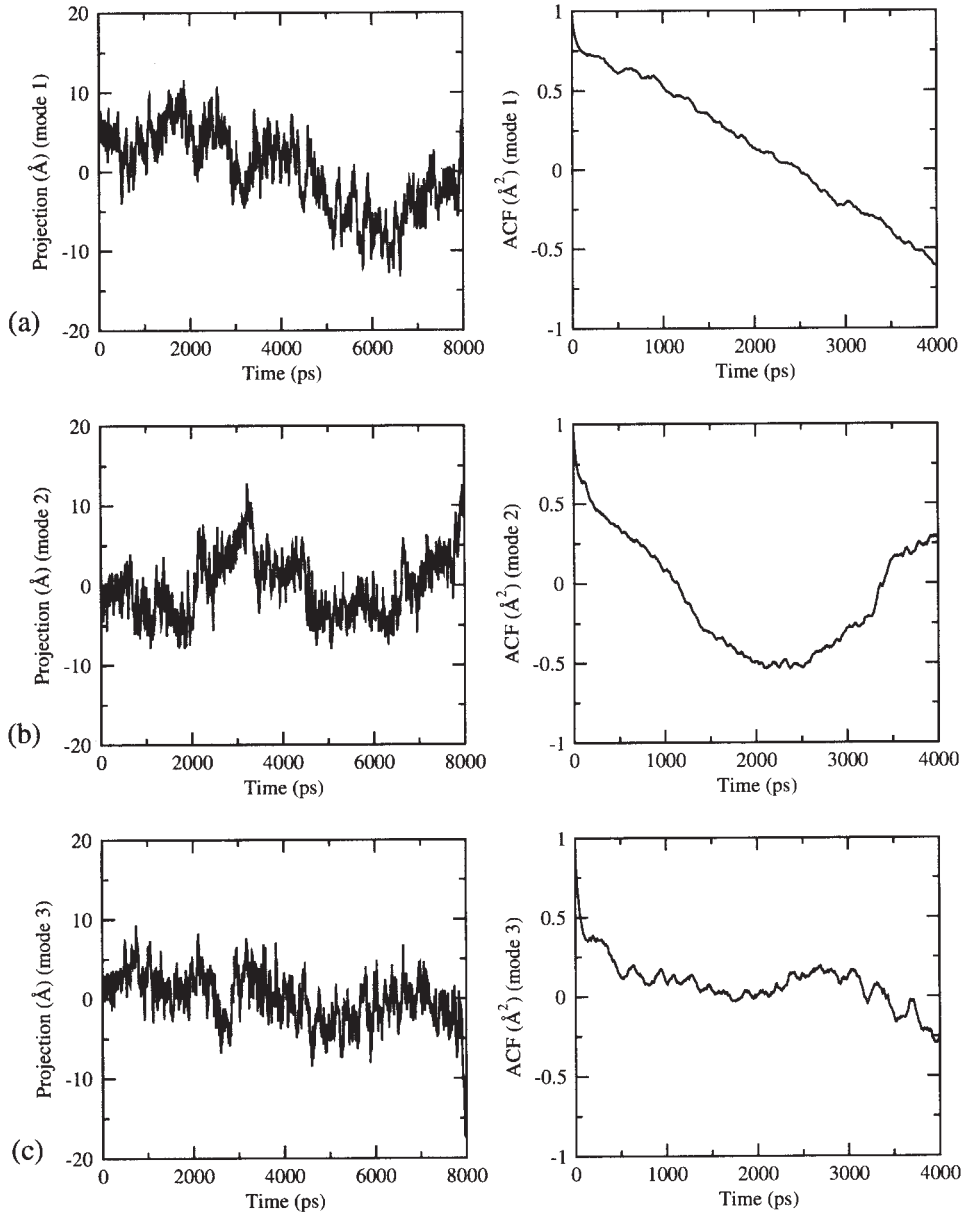

Fig. 7. Projections of the trajectory along the first three PCA modes (left) and their autocorrelation functions (right). (**a**) The first PCA mode, (**b**) the second PCA mode, and (**c**) the third PCA mode.

under sampled global conformational changes that affect the local dynamics, or to noise. We acknowledge that the entire subspace of the first 10 to 15 PCA modes is more robust between time windows[39] compared to the individual mode overlaps shown in Figure 6(a). In LFA, we take advantage of this because the subspace of all localized modes is identical to the standard subspace of the PCA modes used (Eq. 10). This is one of the reasons why most of the LFA modes are well converged even though individual PCA modes are not.

Relaxation Times

To illustrate the actual dynamics observed along individual modes, it is customary to project the MD trajectory along the modes. In the above terminology, we are interested in the outputs of both PCA and LFA representations. The PCA outputs A_r (Eq. 3) have length units and correspond to the deformation of a structure along a mode, whereas the LFA outputs $O(i)$ (Eq. 6) are dimensionless, raising the question about their physical meaning. The LFA outputs preserve all information of the PCA outputs, which are decorrelated and weighted by eigenvalue. $\langle A_r A_q \rangle = \lambda_r \delta_{rq}$. The factor $1/\sqrt{\lambda_r}$ then normalizes the PCA outputs to unity, thus different A_r can be mixed by LFA (Eq. 8). So the LFA outputs $\{O_i\}$, which are decorrelated only in the asymptotic limit $n \to 3N$, give the contribution of each DOF i to the protein dynamics in the low-dimensional subspace $n \ll 3N$.

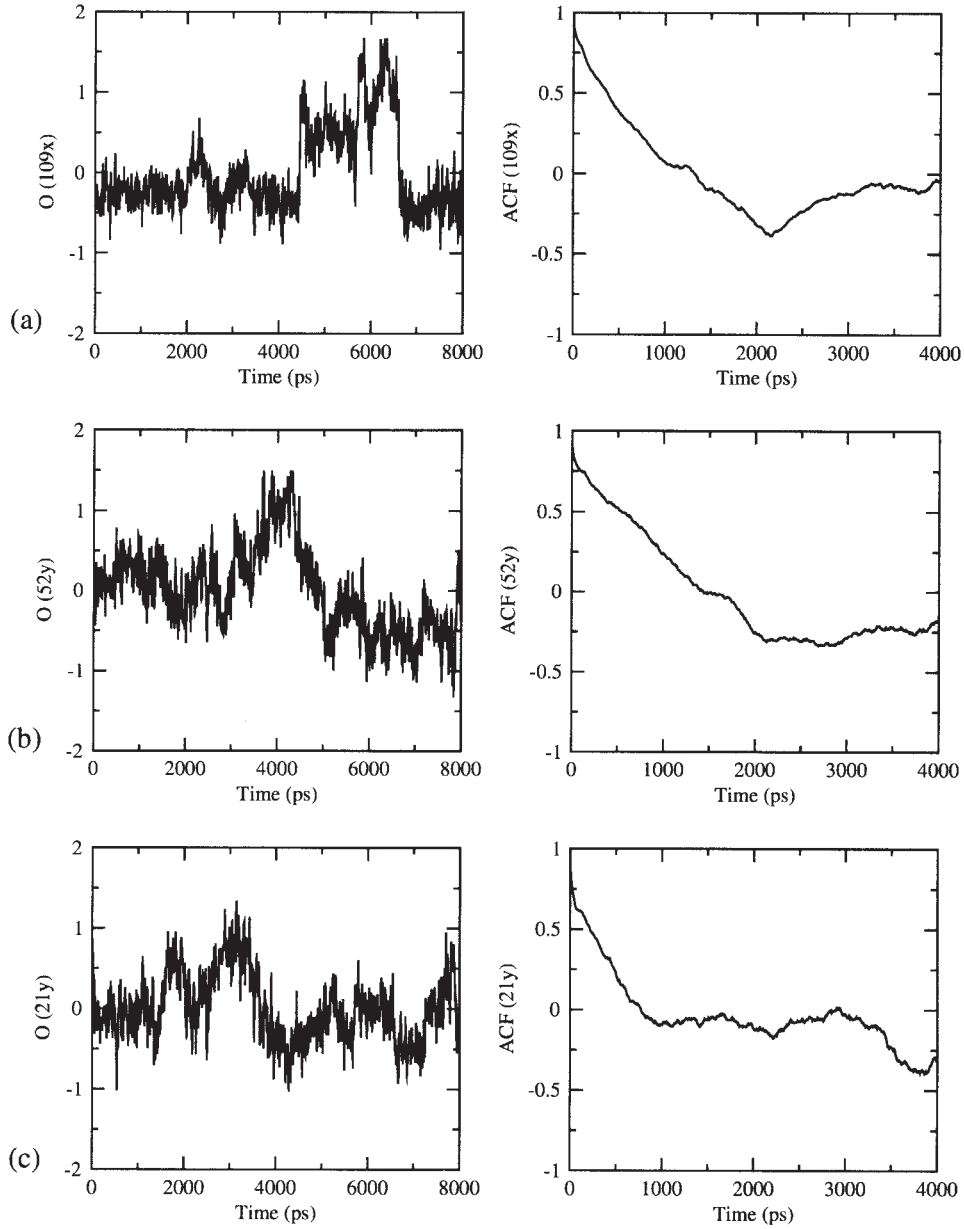

Fig. 8. The output functions of local features (left), and their autocorrelation functions (right). The first 15 PCA modes were used for LFA ($n = 15$ in Eq. 8). Each local feature is represented by a seed atom, and the seed index selected by the sparsification algorithm: (**a**) *x*-component of $C_\alpha - 109$, (**b**) *y*-component of $C_\alpha - 52$, and (**c**) *y*-component of $C_\alpha - 21$.

To compare the performance of PCA and LFA, we first plotted the projections of the trajectory onto the three largest eigenvalue PCA modes and calculated their autocorrelation functions (Fig. 7). The results show that the projections along the low-frequency modes exhibit relaxation times which are longer than the sampling time window, in agreement with Balsera and colleagues.[19] For the dynamics to converge, the autocorrelation function should decay to 0 within the sampling time. This is not observed because such global collective motions are under sampled by MD.

Figure 8 illustrates output functions of three local features ($C_\alpha - 109$, $C_\alpha - 52$, and $C_\alpha - 21$) and their autocorrelation functions. Although the first two LFA modes have relaxed better than their global PCA counterparts, the figure shows that the LFA modes still suffer from long relaxation times. Because the LFA outputs depend on the low-frequency subspace from PCA, it can not be expected that they solve the MD sampling problem simply by virtue of a different statistical analysis. For example, there is a large fluctuation of the local feature corresponding to $C_\alpha - 109$ in time window **II** (from about 6.5 to 8.5 ns, which is also visible in Figure 4(b,c) (red). This rare event corresponding to a transient melting of an

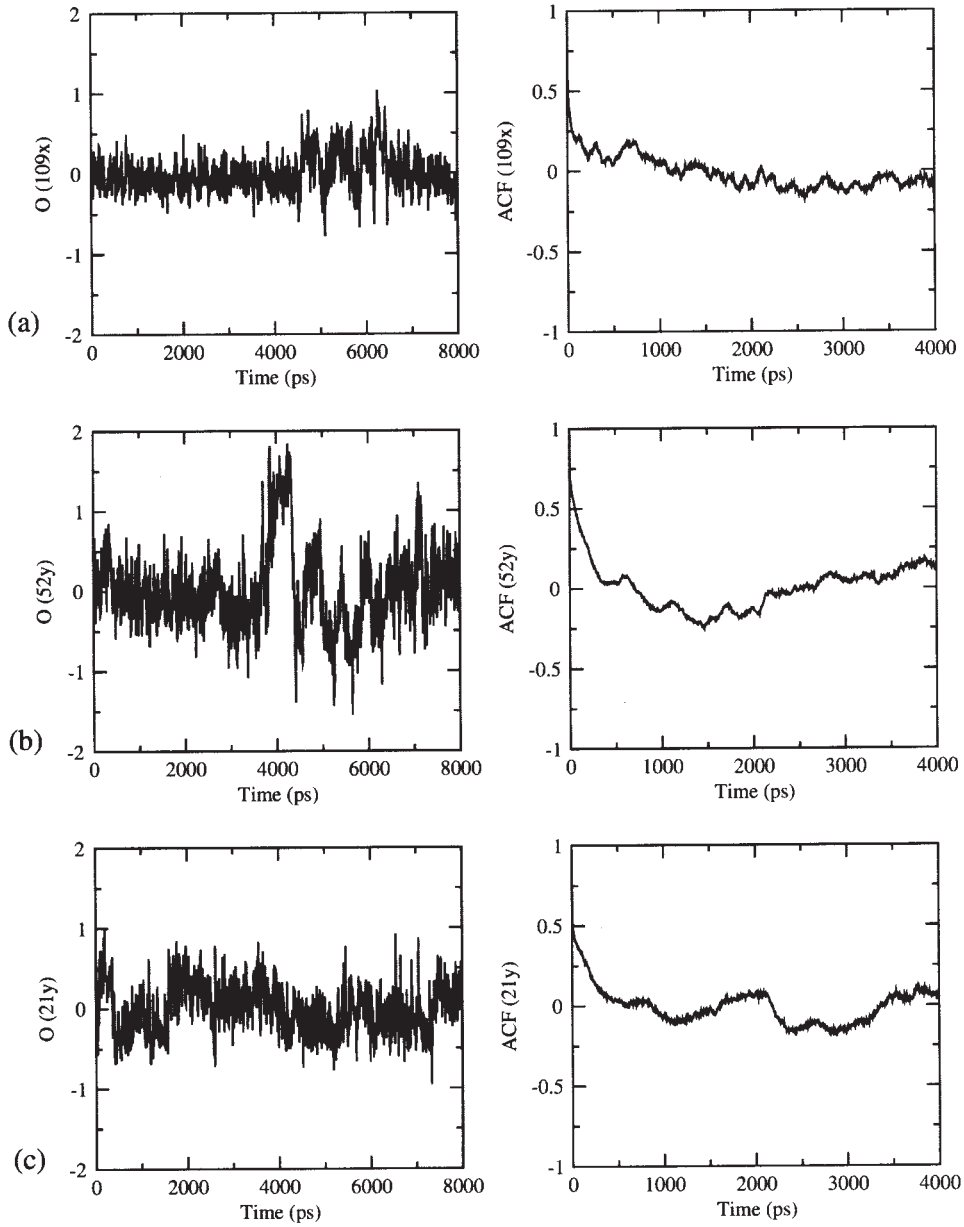

Fig. 9. Same as Figure 8, except that we used a *P*-weighted least-squares fitting of the dynamic domains. In the weighting, negative values of the *P* matrix were set to 0.

α-helix prohibits short relaxation times of the corresponding output.

The most significant reduction in relaxation times can be achieved by focusing on the intrinsic motions of the LFA domains. In Cartesian space, the results of PCA and LFA rely on what reference is used for fitting of the trajectory frames. In the above, we least-squares fitted the structures by all C_α atoms, reflecting the global scope of the PCA modes. However, the assignment of local dynamic domains enables us now to take advantage of the local reference frames of the LFA modes. After a *P*-weighted least-squared fitting, we recalculated the local feature outputs from Equation 8. The results are shown in Figure 9. It can be noticed that the decay of the autocorrelation is now clearly observed in contrast to the globally fitted PCA and LFA modes. Because we eliminated the interdomain motions (which exhibit slower relaxation times), the remaining intradomain relaxation times are of the order of 1 ns, which enables the internal motion to be sampled within the 8 ns simulation time frame.

CONCLUSIONS

This article represents the first application of LFA to the study of protein dynamics. The algorithm enables a segmentation of the system into local dynamic domains. In the model system T4 lysozyme, these local features correspond to the most flexible parts in the protein, and they can be

related to functional domain motions. The overall functional motion can be well described by only a few of the local features.

LFA is a promising tool to study protein dynamics in a reduced dimensional space. One of the major limitations of the traditional PCA-based statistics is the global support of the output basis functions. This effect is due to the forced orthogonalization of the successive modes. In LFA, we construct a local topographic representation of objects in terms of local features from the global PCA modes. The major advantages of LFA are (1) the reproducibility of the modes [different sampling time windows exhibit nearly the same local features, see Figure 6(b) and Table I], and (2) the fast relaxation times of intradomain motions when trajectory frames are aligned by individual domains (Fig. 9).

Does LFA solve the MD sampling problem? As a statistical tool LFA can be used for trajectory analysis but it does not by itself enhance sampling of rare or slow events that are outside of the MD sampling window. However, the method presents a robust way to isolate individual modes that are under sampled, something that is not possible with PCA: Because most LFA modes are converged, one can identify in the overlap matrix the small subset of modes that are affected by noise or by a coupling to an under sampled large-scale motion in the biomolecule.

Do large proteins or macromolecular assemblies really move as statistically independent dynamic domains? Unlike PCA, the *P*-weighted LFA (Fig. 9) does not attempt to estimate a coherence of motion across large distances in biomolecules. Because the sampling of such interdomain coherence is out of reach for short MD simulations, any statistical technique would risk to overestimate such a coherence in an under sampling situation. As a statistical model the *P*-weighted LFA is focused mainly on the sampling of the well-converged intradomain motion and therefore it is better adapted to the short MD dynamics. However, LFA does not rule out that long range interactions exist on much longer time scales.

How many local features are needed to describe the functional dynamics in the protein and how are the dynamic domains defined? We may chose the first n PCA modes that contribute to a certain percentage of the overall motion in the protein. In our system T4 lysozyme, the first 15 PCA modes out of 486 (about 3%) contribute to more than 70% of the total fluctuation. Also, in the sparsification algorithm, we can keep adding seed atoms until the reconstruction error (Eq. 14) is below an acceptable value. In our case, the reconstruction error decreases about 60% after 5 seed atoms out of 15 are selected. These five seed atoms are the C-terminal atom, $C_\alpha - 109$, $C_\alpha - 52$, $C_\alpha - 21$, and the N-terminal atom, respectively [Fig. 1(d)]. However, any n atoms can be used in principle to reconstruct the $O(i)$ without error.

One possible improvement of the sparsification would be a simultaneous instead of a sequential optimization of the seed atoms involving a criterion that directly minimizes the correlation between the dynamic domains. Currently, we define the boundary of a dynamic domain by the contiguous atoms that have positive correlations with the seed atom in terms of LFA theory. A threshold above the background noise level of the correlation may be an alternative. These issues will be subject of future research. Overall, our initial work presented here demonstrates that LFA shows much promise for many applications in prediction, sampling and classification of large-scale macromolecular structure and dynamics.

ACKNOWLEDGMENTS

We are grateful to Dr. Danny C. Sorensen at Rice University for helpful discussions. Z. Z. was supported by the Keck Center Pharmacoinformatics Training Program of the Gulf Coast Consortia.

REFERENCES

1. Karplus M. Molecular dynamics: applications to proteins. In: J.-L. Rivail, editor. Modelling of molecular structures and properties, volume 71, Studies in physical and theoretical chemistry. Amsterdam: Elsevier Science Publishers; 1990. p 427–461.
2. Brooks CL III, Karplus M, Pettitt BM. Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics, volume LXXI of Advances in Chemical Physics. New York: John Wiley & Sons, 1988.
3. Hernández G, Jenney FE Jr., Adams MWW, LeMaster DM. Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. Proc Natl Acad Sci U S A 2000; 97:3166–3170.
4. Falke JJ. A moving story. Science 2002;295:1480–1481.
5. Horiuchi T, Go N. Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. Proteins 1991;10:106–116.
6. Kitao A, Hirata F, Go N. The effects of solvent on the conformation and the collective motions of proteins: normal mode analysis and molecular dynamics simulations of Melittin in water and in vacuum. Chem Phys 1991;158:447–472.
7. Amadei A, Linnsen ABM, Berendsen HJC. Essential dynamics of proteins. Protein 1993;17:412–425.
8. van Aalten MF, Amadei A, Linssen ABM, Eijsink VGH, Vriend G, Berendsen HJC. The essential dynamics of Thermolysin: conformation of the hinge-bending motion and comparison of simulations in vacuum and water. Proteins 1995;22:45–54.
9. García AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett 1992;68:2696–2699.
10. Kitao A, Go N. Investigating protein dynamics in collective coordinate space. Curr Opinion Struct Biol 1999;9:164–169.
11. Berendsen HJC, Hayward S. Collective protein dynamics in relation to function. Curr Opinion Struct Biol 2000;10:165–169.
12. Brooks BR, Janezic D, and Karplus M. Harmonic analysis of large systems I. Methodology. J Comput Chem 1995;16:1522–1542.
13. Case DA. Normal mode analysis of protein dynamics. Curr Opinion Struct Biol 1994;4:285–290.
14. Karhunen K. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. Ann Acad Sci Fennicae Ser. A137, 1947.
15. Karplus M, Jushick JN. Method for estimating the configurational entropy of macro-molecules. Macromolecules 1981;14:325–332.
16. Levy RM, Srinivasan AR, Olson WK, McCammon JA. Quasiharmonic method for studying very low frequency modes in proteins. Biopolymers 1984;23:1099–1112.
17. Kitao A, Hayward S, Go N. Energy landscape of a native protein: jumping-among-minima model. Proteins 1998;33:496–517.
18. Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips GN. A sampling problem in molecular dynamics simulations of macromolecules. Proc Natl Acad Sci U S A 1995;92:3288–3292.
19. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal component analysis and long time protein dynamics. J Phys Chem 1996;100(7):2567–2572.
20. Tama F, Sanejouand Y.-H. Conformational change of proteins arising from normal mode calculations. Protein Eng 2001;14:1–6.
21. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–791.
22. Paatero P. Least squares formulation of robust non-negative factor analysis. Chemometrics Intell Lab Sys 1997;37:23–35.

23. Penev PS, Atick JJ. Local feature analysis: a general statistical theory for object representation. Network: Computation in Neural Systems 1996;7:477–500.
24. Janezic D, Venable RM, Karplus M. Harmonic analysis of large systems III. Comparison with molecular dynamics. J. Comp Chem 1995;16:1554–1568.
25. Li Z, Atick J. Towards a theory of the striate cortex. Neural Comp 1994;6:127–146.
26. Berendsen HJC, van der Spoel D, van Drunen R. Gromacs: A message-passing parallel molecular dynamics implementation. Comput Phys Commun 1995;91:43–56.
27. Lindahl E, Hess B, van der Spoel D. Gromacs 3.0: A package for molecular simulation and trajectory analysis. J Mol Model 2001;7: 306–317.
28. Anderson WF, Grutter MG, Remington SJ, Weaver LH, Matthews BW. Crystallographic determination of the mode of binding of oligosaccharides to T4 bacteriophage lysozyme: implications for the mechanism of catalysis. J Mol Biol 1981;147:523–543.
29. Faber HR, Matthews BW. A mutant T4 lysozyme displays five different crystal conformations. Nature 1990;348:263–266.
30. Zhang XJ, Wozniak JA, Matthews BW. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J Mol Biol 1995;250:527–552.
31. Weaver LH, Matthews BW. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. J Mol Biol 1987;193:189–199.
32. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J. Interaction models for water in relation to protein hydration. In: B. Pullman, editor. Intermolecular forces. Dordrecht, Netherlands; Reidel 1981. p. 331–342.
33. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR, Molecular dynamics with coupling to an external bath. J Chem Phys, 1984;81(8):3684–3690.
34. Hess B, Bekker H, Berendsen HJC, Fraaije GJEM. A linear constraint solver for molecular simulations. J Comput Chem 1997;18:1463–1472.
35. Essman U, Perela L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh ewald method. J Chem Phys 1995;103: 8577–8592.
36. Humphrey WF, Dalke A, Schulten K. VMD - Visual Molecular Dynamics. J Mol Graphics 1996;14:33–38.
37. de Groot BL, Hayward S, van Aalten DMF, Amadei A, Berendsen HJC. Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. Proteins 1998;31:116–127.
38. Zhang Z, Shi Y, Liu H. Molecular dynamics simulations of peptides and proteins with amplified collective motions. Biophys J 2003;84:3583–3593.
39. Amadei A, de Groot BL, Ceruso MA, Di Nola A, Berendsen HJC. A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells. Proteins 1999;35:283–292.