Coarse-Graining Protein Structures With Local Multivariate Features from Molecular Dynamics

Zhiyong Zhang[†] and Willy Wriggers*[,‡]

School of Health Information Sciences, University of Texas Health Science Center at Houston, 7000 Fannin ST, Suite 600, Houston, Texas 77030

Received: July 16, 2008; Revised Manuscript Received: September 3, 2008

A multivariate statistical theory, local feature analysis (LFA), extracts functionally relevant domains from molecular dynamics (MD) trajectories. The LFA representations, like those of principal component analysis (PCA), are low dimensional and provide a reduced basis set for collective motions of simulated proteins, but the local features are sparsely distributed and spatially localized, in contrast to global PCA modes. One key problem in the assignment of local features is the coarse-graining of redundant LFA output functions by means of seed atoms. One can solve the combinatorial problem by adding seed atoms one after another to a growing set, minimizing a reconstruction error at each addition. This allows for an efficient implementation, but the sequential algorithm does not guarantee the optimal mutual correlation of the sequentially assigned features. Here, we present a novel coarse-graining algorithm for proteins that directly minimizes the mutual correlation of seed atoms by Monte Carlo (MC) simulations. Tests on MD trajectories of two biological systems, bacteriophage T4 lysozyme and myosin II motor domain S1, demonstrate that the new algorithm provides statistically reproducible results and describes functionally relevant dynamics. The well-known undersampling of large-scale motion by short MD simulations is apparent also in our model, but the new coarse-graining offers a major advantage over PCA; converged features are invariant across multiple windows of the trajectory, dividing the protein into converged regions and a smaller number of localized, undersampled regions. In addition to its use in structure classification, the proposed coarse-graining thus provides a localized measure of MD sampling efficiency.

Introduction

Molecular dynamics (MD) simulations have become a popular method for examining dynamical properties of large biological systems.[1−4] These systems typically consist of thousands of degrees of freedom (DOF), rendering all but the most trivial quantitative analysis of the trajectory prohibitively complex. One interesting problem in MD analysis is the identification of collective modes in the configurational space corresponding to the dominant global motions observed in the high-dimensional trajectory.

It is useful to attempt to reduce the dimensionality of the problem to facilitate the assignment of functionally relevant motions. This simplification is possible due to the probability distributions of the MD trajectory along principal modes determined by quasi-harmonic analysis. Principal component analysis (PCA) of high-dimensional time series data employs a diagonalization of the symmetric covariance matrix.[5,6] The resulting eigenvectors are the principal modes describing the motion of the system. The corresponding eigenvalues, sorted in decreasing order, provide a measure of the observed magnitudes of these motions. For all but a few n of the largest eigenvalues, the projections of the trajectory onto the eigenvectors are unimodal and can be well approximated by Gaussian functions. Therefore, along these "inessential" coordinates in $3N - n$ dimensions, the corresponding free-energy landscape is simply harmonic, and the dynamics can be described as oscillations about the equilibrium positions. The remaining orthogonal subspace, however, contains the "essential" modes, which correspond to a more complex dynamics that cannot be described by harmonic potential wells. For most systems, the dimension n of this essential space is very small, $n \ll 3N$.

Due to their appeal, PCA-based dynamics techniques[7,8] have been applied for some time to sample the conformational space.[6,9,10] However, it became apparent in the mid 1990s that there are limitations to PCA conferred by the MD sampling problem. Relaxation times of correlations in real proteins are on the order of milliseconds or longer,[11,12] whereas MD simulations are currently restricted to the time scale of nanoseconds to microseconds, much shorter than the time scales of many important biological processes. Therefore, the correlations in the essential subspace are undersampled by nanosecond MD simulations,[13] and PCA modes from short MD trajectories are intrinsically unreliable.[14] Also, the global extent of individual PCA modes is problematic because of their forced orthogonalization, which creates complex dependencies among the eigenvectors.[15]

One solution to the sampling problem is longer simulation times,[6,16] but it is difficult to know "how long is long enough". Therefore, we have been seeking an alternative multivariate statistical theory that describes dynamic features locally to avoid the sampling and orthogonalization problems. Penev and Atick originally developed a statistical technique, termed local feature analysis (LFA), for face recognition and image classification.[17] Starting from PCA, the n-dimensional essential subspace is

* To whom correspondence should be addressed. Tel.: (212) 403-8131. Fax: (646) 873-2131. E-mail: wriggers@biomachina.org.

† Current address: Center for Biophysical Modeling and Simulation and Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, UT 84112-0850.

‡ Current address: D.E. Shaw Research, 39th Floor, 120 West 45th Street, New York, NY 10036.

10.1021/jp806291p CCC: $40.75
Published on Web 10/15/2008

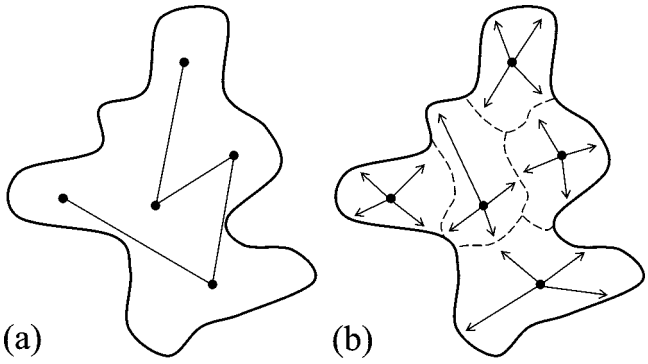

Figure 1. Schematic diagram illustrating CGLC. (a) Coarse-graining by minimizing mutual correlation among neighboring seed atoms in a linear chain following the polypeptide fold. The seed correlation is given by eq 12. (b) Assignment of dynamic domains which are contiguous in sequence and whose motion is highly correlated with the seed atom.

expanded to localized features in the 3*N*-dimensional coordinates. In ref 15, we devised an algorithm termed "sparsification" to extract *n* local features from the 3*N*-dimensional LFA outputs. The sparsification (or coarse-graining) reduces the LFA outputs to a small subset of seed atoms that correspond to the mobile parts of the system. The surrounding "dynamic domains" are then defined as the regions of high output correlation with the seed atoms.

In the present work, we aim to enhance the coarse-graining such that seeds are purposefully distributed in a nonredundant manner and to achieve full coverage of the protein. In the following section, we introduce a novel coarse-graining technique termed coarse-graining of linear chains (CGLC) specifically for protein architectures. CGLC performs a simultaneous optimization of the seed atoms by minimizing the mutual correlation between the local features (Figure 1a). Beginning with a random set of seed atoms, we perform a Monte Carlo simulation for this purpose. After the seeds are optimized, local dynamic domains are assigned, as before, by identifying regions of positive correlation surrounding the corresponding seeds (Figure 1b). We describe the application to two biologically relevant systems, T4 lysozyme (T4L), and myosin II motor domain S1 (MYO). The tests below show that the new algorithm is able to provide reproducible coarse representations for regions whose dynamics is well converged.

In the following, we provide an abridged review of LFA theory, followed by a description of the available coarse-graining algorithms. Also, computational details for the MD simulations are given. Subsequently, we evaluate the performance and discuss the advantages of CGLC. The effect of the "MD sampling problem" on the reproducibility of seeds and local dynamic domains is shown, and applications in convergence analysis are discussed. Finally, we give concluding remarks on the prediction, sampling, and classification of large-scale macromolecular dynamics and provide a practical implementation road map.

Theory and Methods

For a system consisting of *N* atoms (it is useful here to focus only on the C_α atoms of a protein), the internal motion (after eliminating the overall translational and rotational motion) is described by a trajectory $x(t)$, where x is a 3*N*-dimensional column vector of the C_α atomic coordinates, $\{x_1, x_2, ..., x_{3N}\}$. 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) $\Psi_r(i)$, $r = 1,..., 3N$ and corresponding eigenvalues λ_r

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

We sort λ_r in a decreasing order up to a small, user-defined, number n ($n \ll 3N$) of modes that are sufficient to describe the dominant dynamics (n is thus the dimension of the essential subspace).

Overview of LFA Theory. In ref 15, we derived the theory of LFA for biomolecular dynamics that can be summarized in compact form as follows. We define a family of matrices

$$K^{(m)}(i,j) = \sum_{r=1}^{n} \Psi_r(i) \left(\frac{1}{\sqrt{\lambda_r}} \right)^m \Psi_r(j) \quad (3)$$

where the cases $m = 1, 0, -1, -2$ are of specific importance:

• $K^{(1)}(i,j) \equiv K(i,j)$ is the so-called kernel that projects the trajectory onto individual modes. Using the kernel, we define the so-called output of the LFA projection

$$O(i) \equiv \sum_{j=1}^{3N} K(i,j) \Delta x_j = \sum_{r=1}^{n} \frac{A_r}{\sqrt{\lambda_r}} \Psi_r(i) \quad (4)$$

In the more familiar PCA formalism, the kernel is $K_r(i) = \Psi_r(i)$, and the output $A_r = \sum_{j=1}^{3N} K_r(j) \Delta x_j$ is the projection of atomic motions onto the PCA mode Ψ_r. The factor $1/\sqrt{\lambda_r}$ in LFA normalizes the PCA output A_r. Thus, different A_r can be mixed in eq 4. The LFA outputs preserve all of the information present in the PCA outputs, but they are localized (indexed by atoms i not modes r).

• $K^{(0)} \equiv P(i,j)$ is the residual output correlation. Since the 3*N* outputs $O(i)$ (eq 4) are derived from $n \ll 3N$ principal modes, the LFA outputs are not fully decorrelated; instead, one can show that

$$\langle O(i) O(j) \rangle = P(i,j) \quad (5)$$

The LFA outputs $O(i)$ become orthogonal and normalized to unity ($P(i,j) \to \delta(i,j)$) only in the limit $n \to 3N$.

• $K^{(-1)}$ is the reconstructor (inverse LFA kernel)

$$\Delta x_i^{\text{rec}} \equiv \sum_{j=1}^{3N} K^{(-1)}(i,j) O(j) \quad (6)$$

where the reconstructed positions Δx_i^{rec} approximate the original positions Δx_i in a least-squares sense.

• The asymptotic limit of $K^{(-2)}$ is the familiar covariance matrix: $K^{(-2)} \to C(i,j)$ for $n \to 3N$.

The matrices $K = K^{(1)}$ and $P = K^{(0)}$ are central to LFA. K is by definition the projection operator onto a local feature. Also, it is straightforward to show that

$$\sum_{j=1}^{3N} P(i,j)\Delta x_j = \Delta x_i^{\mathrm{rec}} \tag{7}$$

This means that P serves a dual role both as the correlation of the LFA outputs (see above) and as the projection operator onto the low-frequency subspace spanned by n PCA modes.

Coarse-Graining. LFA theory replaces the n global PCA modes with $3N$ local LFA output functions $O(i)$ (eq 4). Although locality is achieved, it comes initially at the price of expanding again to the full number $3N \gg n$ of DOF. Since these LFA outputs are highly redundant, an additional coarse-graining step must be applied to the LFA output space, recovering a low-dimensional representation. In the case of biomolecular structures, we force the sparse outputs to be distributed among n distinct atoms, the so-called seed atoms. The number n of seed atoms is chosen to be identical to the dimension n of the essential PCA subspace.[15]

The problem of coarse-graining can then be formulated as an optimization problem where n indistinguishable seed atoms are located among N atoms in the system. The computational complexity of the optimization is given by the binomial coefficient $\binom{N}{n} = N!/n!(N-n)!$. The focus of this work is on finding a nonredundant and reproducible distribution among the $\binom{N}{n}$ possible seed configurations that provides a functionally meaningful coverage of the protein.

Earlier "Sparsification" Algorithm. In our earlier LFA paper,[15] we adapted a sequential algorithm[17] termed "sparsification". We approximated the $3N$ outputs $O(i)$ with only a small subset $\mathscr{M}$ of outputs that correspond to the strongest local features. The other $O(i)$ can then be reasonably well predicted by the subset

$$O^{\mathrm{rec}}(i) = \sum_{m=1}^{|\mathscr{M}|} a_m(i) O(i_m) \tag{8}$$

where $a_m(i)$ is the optimal linear prediction coefficient, which is defined to minimize the average reconstruction mean square error on the $O(i)$

$$E^{\mathrm{rec}} = \langle \| O^{\mathrm{err}}(i) \|^2 \rangle \equiv \langle \| O(i) - O^{\mathrm{rec}}(i) \|^2 \rangle \tag{9}$$

The $a_m(i)$ is determined by

$$a_m(i) = \sum_{l=1}^{|\mathscr{M}|} P(i, i_l)(P'^{-1})_{lm} \tag{10}$$

where P'^{-1} is the inverse of the submatrix P' from P, $P'_{lm} \equiv P(i_l, i_m)$. Beginning with an empty set $\mathscr{M}$, we added at each step $m+1$ the one seed index (corresponding to a unique new seed atom), which exhibits the maximum reconstruction error $O^{\mathrm{err}}(i_{m+1})$. We kept adding seed indices to $\mathscr{M}$ until n of them were chosen.

Since the criterion E^{rec} used in eq 9 vanishes toward the end, the minimization of E^{rec} exerts stronger constraints on the initial members of $\mathscr{M}$. To promote an even distribution of seeds, we defined an additional neighborhood exclusion number, that is, the newly added seed atom and its nearest neighbors in primary sequence (within the exclusion number) were forced to be distinct from atoms corresponding to previously found indices. This extra parameter improved the robustness of the earlier coarse-graining and ensured that seed atoms would be evenly distributed. However, this additional parameter was not rooted in statistical theory.

We introduce in the following the novel CGLC criterion based on statistics alone that avoids the problems of the sequential algorithm and takes advantage of the linear chain architecture specific to proteins. In contrast to CGLC, we denote the previous "sparsification" algorithms as SPA-0 (using only self-exclusion of found seed atoms) and SPA-k (excluding a number $k > 0$ of neighboring seeds in C_α representations of proteins).

Monte Carlo Coarse-Graining of Linear Chains. In the new algorithm CGLC, we deal with the seed atoms directly instead of the DOF (seed indices). Considering an atom h, its LFA output $\vec{O}_h$ has three components, $O(h_d)$, $d = 1$, 2, or 3. The correlation between the atom h and another atom k (with LFA output $\vec{O}_k$) is

$$\langle \vec{O}_h \cdot \vec{O}_k \rangle = \sum_{d=1}^{3} \langle O(h_d) O(k_d) \rangle \equiv \sum_{d=1}^{3} P(h_d, k_d) \tag{11}$$

Therefore, the correlation between any two atoms is the trace of a 3×3 submatrix within $P(i,j)$. In empirical tests of LFA, we found that a seed atom is typically surrounded by a positively correlated dynamic domain, followed at larger distances by less correlated regions with smaller and noisy correlations.[5,15] For a nonredundant placement of n seed atoms into the set $\mathscr{M}$, the output correlations between the seeds h and neighbors $h+1$ must be small to avoid an overlap between the immediately surrounding positive regions. Considering only neighbors in the polypeptide chain, it would be unnecessary to optimize the output correlations between the seed atom h and other distant seed atoms k ($k > h+1$). Thus, we define a new optimization criterion, termed "linear chain seed correlation", that penalizes overlap of positive neighbor correlations

$$E^{\mathrm{lcsc}} = \sum_{h=1}^{n-1} \langle \vec{O}_h \cdot \vec{O}_{h+1} \rangle \tag{12}$$

The linear chain seed correlation is the sum of all of the neighboring pairwise output correlations in the set $\mathscr{M}$. Starting with any initial set $\mathscr{M}$, we minimize eq 12 to obtain a nonredundant distribution of seeds. This simultaneous optimization is different from the old algorithm, in which $\mathscr{M}$ was filled incrementally. For the optimization of eq 12, we use the well-known Metropolis Monte Carlo (MC)[18] algorithm.

In MC, a new set $\mathscr{M}$ is accepted or rejected based on the Metropolis criterion. Downhill moves are always accepted; if the new set exhibits a lower seed correlation E_1^{lcsc} than its predecessor (E_0^{lcsc}) according to eq 12, then the new set is accepted as the starting point for the next iteration. Uphill moves ($\Delta E^{\mathrm{lcsc}} = E_1^{\mathrm{lcsc}} - E_0^{\mathrm{lcsc}} > 0$) are accepted only with a probability of $\exp(-\Delta E^{\mathrm{lcsc}}/T_{\mathrm{m}})$, where T_{m} is a temperature-like parameter. In practice, the exact parameter value is not critical. Multiple MC simulations are carried out with different T_{m} values, and the maximum T_{m} is chosen to be about 10% of the peak output correlation. Uphill moves ($\Delta E^{\mathrm{lcsc}} > 0$) then occur with reasonable acceptance frequency.

In summary, our MC algorithm can be described as follows: (1) calculate the correlation (eq 12) of the initial seed atom set; (2) change one seed atom randomly by drawing from the nonoccupied atoms; (3) calculate the correlation of the new set; (4) accept or reject the new seed atom set according to the

Metropolis criterion; and (5) repeat the procedure for a number of predefined steps (here, $n \cdot 1000$). We will then pick the lowest-correlation set after MC minimization.

Following the stochastic global search, we perform also (see Results and Discussion) a final steepest descent minimization that preserves the approximate location of seeds but allows small moves to find the nearest local minimum of eq 12. In each iteration, one chosen seed atom is changed to its $+1$ or -1 position in primary sequence that offers the most favorable minimization among all of the possible position changes. This step is repeated until convergence.

Molecular Dynamics Simulations. Details of the MD simulation of T4L were described in ref 15. The MD simulation of MYO was carried out using the GROMOS96 simulation package with a united-atom parameter set 43A1.[19,20] The initial structure was taken from the supplementary structure "motor_domain.pdb" published by Holmes et al.[21] The simulated myosin system consists of a main chain that is interrupted at three missing loops (residue numbers 4-843[22]), a regulatory light chain (RLC, residue numbers 859-1008[21]), and an essential light chain (ELC, residue numbers 1016-1160[21]). The protein was placed in a rectangular box such that the minimal distance between the solute and the box boundary was 0.8 nm. SPC water molecules were added from an equilibrated cubic box containing 216 water molecules.[23] Some of the added water molecules were removed so that no water oxygen atom was closer than 0.23 nm to a non-hydrogen atom of the protein or another water oxygen atom. The system, protein and water, was initially energy-minimized using the steepest descent method, until the absolute value of the change in the total potential energy was smaller than 0.1 kJ mol^{-1}. Thirty-two Na^+ ions were added to compensate the net negative charges on the protein by replacing water molecules with the lowest electrostatic potential. The system (protein, ions, and water molecules) was energy-minimized again using the steepest descent method. The final system consisted of myosin (1099 residues and 11216 atoms), 32 Na^+ ions, and 57650 water molecules, leading to a total size of 184 198 atoms. A 100 ps simulation with a positional restraints force constant of 2500 kJ mol^{-1} nm^{-2} was performed at 300 K. Then, a 4 ns production run was performed.

The Verlet integration scheme (leapfrog)[24] with an isothermal−isobaric simulation algorithm[25] and a 2 fs time step was used. Solute and solvent was coupled separately to a temperature bath of 300 K, with a coupling time of 0.1 ps. The pressure was adjusted to 1 atm with a coupling time of 0.5 ps, and the isothermal compressibility was chosen to be 4.575×10^{-4} (kJ mol^{-1} nm^{-3})$^{-1}$. Covalent bonds in the protein were constrained using the SHAKE method with a relative geometric tolerance of 10^{-4}.[26] Long-range forces were treated using twin-range cutoff radii of $R_{cp} = 0.8$ nm for the charge group pair list and $R_{cl} = 1.4$ nm for the longer-range nonbonded interaction. The pair list was updated every 20 fs. Reaction field forces were included,[27] corresponding to a dielectric permittivity of 54 for SPC water.[28]

Results and Discussion

Our main goal is the coarse-graining of LFA seeds; therefore, we describe in the following in detail the performance tests of CGLC. We investigated specifically whether CGLC can overcome the limitations of SPA-k ($k \geq 0$) in terms of redundancy and reproducibility of the found seed atoms. Also, the resulting dynamic domains are visualized, and a brief biological discussion of the functional significance is given for the two systems studied.

Bacteriophage T4 Lysozyme. Local features were extracted from the 2−10 ns "production" part of the T4L trajectory. If we consider four seed atoms ($n = 4$) in T4L, the result is (40, 51, 53, 109) using SPA-0. The seed atoms C_α-51 and C_α-53 are too close, and their output correlation is 0.095. The output correlation between the atoms C_α-40 and C_α-51 is also positive (0.011). The results suggest that the three seed atoms actually represent the same local feature. A less redundant distribution can be achieved with SPA-50, which yields (1, 51, 109, 162). These seeds are now almost evenly distributed in the protein. Besides C_α-51 and C_α-109, the two termini (C_α-1 and C_α-162) are selected because the open ends of the polypeptide chain are typically very flexible. The result of SPA-50 is better than that of SPA-0 in terms of redundancy, but the even distribution was achieved by an artificial constraint, not statistics.

As a preliminary test of CGLC, we started with the seed atom sets resulting from SPA-0 ($E^{lcsc} = 0.042$) and SPA-50 ($E^{lcsc} = -0.135$). As described in Theory and Methods, we minimized the linear chain seed correlation E^{lcsc} (eq 12) systematically. Both examples (Figure 2a,b) converge to the same seed set (3, 51, 108, 162), which is close to the SPA-50 result, reaching a seed correlation of -0.179. Furthermore, we used two extremely uneven distributed start sets (1, 2, 3, 4) and (159, 160, 161, 162), with high seed correlations of 0.099 and 0.059, respectively. Both sets quickly converged to the same solution (Figure 2c,d).

The seed atoms (3, 51, 108, 162) consistently found by CGLC are not redundant as was the case with SPA-0. Also, the even distribution is achieved without any artificial constraint, as was the case with SPA-50. We note that these four seed atoms correspond to the four most flexible regions in T4L (Figure 3a) as measured by the root-mean-square fluctuation (RMSF) of the C_α atoms.

Clearly, if we wish to use only four local features to describe the protein dynamics, these four would be the optimum choice. On the basis of this preliminary test, the linear chain seed correlation (eq 12) appears to be more a more reliable search criterion compared to the reconstruction error (eq 9).

After these promising first tests, we carried out a more systematic investigation of the new CGLC algorithm. For the case of $n = 4$, we used 200 random start sets. For each start set, we performed MC simulations with eight different T_m values (0.03, 0.025, 0.02, 0.015, 0.01, 0.005, 0.002, and 0.001). The maximum temperature parameter equals about 10% of the peak output correlation in the protein, and the other seven T_m values were obtained by decreasing gradually from the maximum T_m. We picked the lowest-correlation set from each of the eight MC minimizations and then used it as the starting point for another round of MC simulations with the eight T_m values. We repeated this procedure until the lowest-correlation set among the eight converged. This was followed by steepest descent minimization (see Theory and Methods).

The results for the 200 test runs are listed in Table 1. In the case of $n = 4$, all 200 initial sets converged to the previously found set (3, 51, 108, 162), in agreement with the results in Figure 2. In summary, the results for $n = 4$ indicate that the seed correlation (eq 12) enables the reproducible identification of four nonredundant local features.

We carried out similar performance tests for the cases of $n = 5$, 6, 7, and 8. Details for these tests with larger n are also given in Table 1. For example, 192 of the 200 initial sets for $n = 5$ converged to (2, 22, 53, 109, 162) with a seed correlation of -0.187. The rms fluctuations (Figure 3a) show that the five

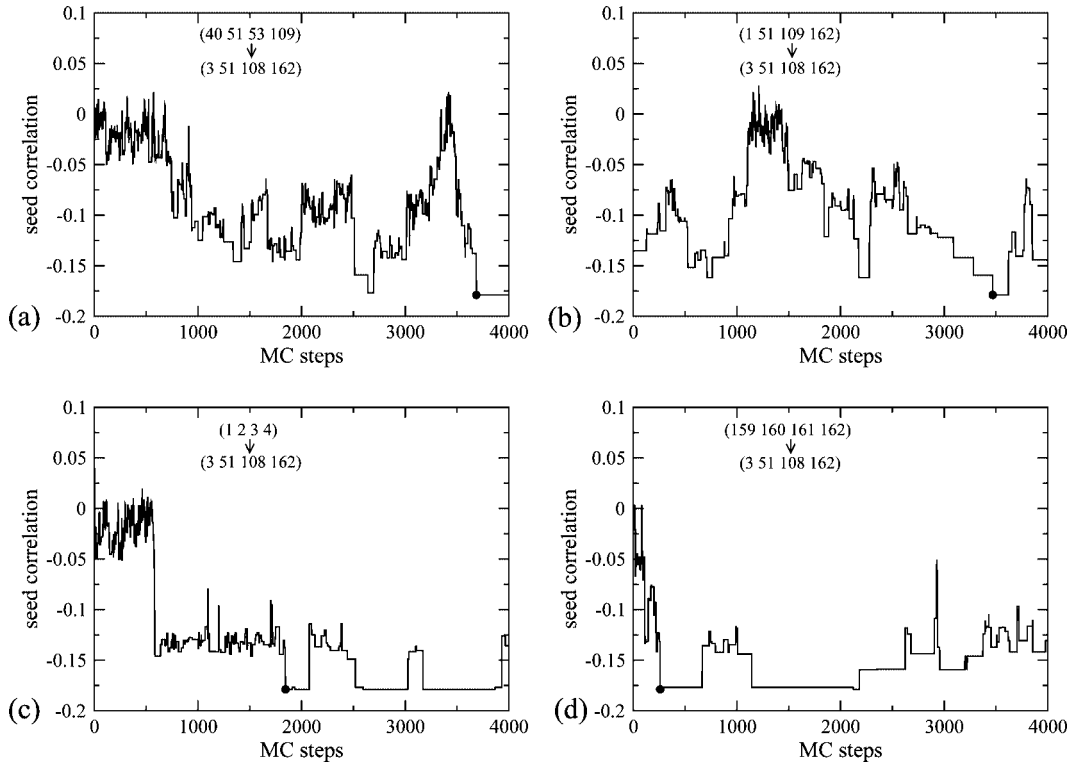

Figure 2. Coarse-graining results of T4L using CGLC for $n = 4$. (a) Starting with the initial seed atom set (40, 51, 53, 109) from SPA-0. (b) Starting with the initial seed atom set (1, 51, 109, 162) from SPA-50. (c) Starting with the initial seed atom set (1, 2, 3, 4). (d) Starting with the initial seed atom set (159, 160, 161, 162). The seed atoms in the lowest-correlation set (3, 51, 108, 162) are indicated by black dots.

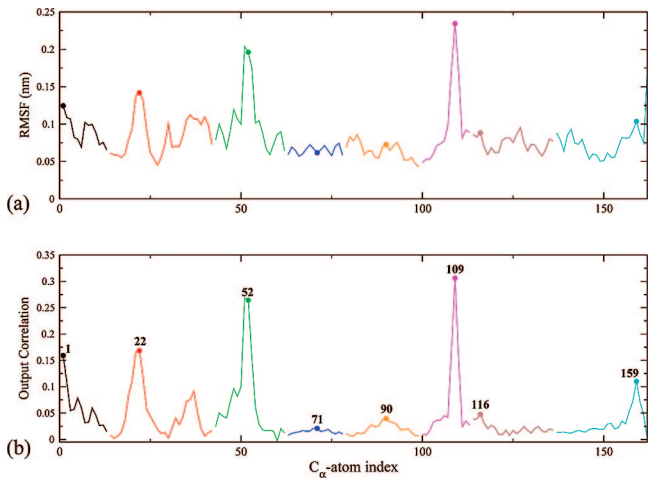

Figure 3. (a) Root-mean-square fluctuations (RMSF) of C_α atoms calculated from the MD simulation of T4L. The RMSF curve is divided into the eight local features (Table 1). C_α-1 (1−13, which means the local feature spans residues 1−13) black; C_α-22 (14−42) red; C_α-52 (43−62) green; C_α-71 (63−78) blue; C_α-90 (79−99) orange; C_α-109 (100−113) magenta; C_α-116 (114−136) brown; and C_α-159 (137−162) cyan. (b) Output correlations (eq 11) between the seed atoms and T4L as a function of residue number in the case of $n = 8$. Only the correlations between one seed atom and the atoms in its corresponding dynamic domain are plotted, which are colored according to the RMSF curve. The eight seed atoms are indicated by dots.

seeds correspond to well-defined regions of high mobility in the structure.

Each dynamic domain is defined as a contiguous and localized set of atoms exhibiting positive output correlation (eq 11) with the corresponding seed atom. The four local features in the case of $n = 4$ cover about 44% of residues in T4L (Table 1). By adding more seed atoms, the coverage is increased until eight local features cover the entire protein (Table 1). We plot the

TABLE 1: T4L Coarse-Graining Using the CGLC Algorithm As a Function of the Number of Local Features, *n*

n	seed atom set	seed correlation[a]	occurrence[b]	coverage (%)[c]
4	(3, 51, 108, 162)	−0.179	200	44
5	(2, 22, 53, 109, 162)	−0.187	192	48
6	(2, 51, 109, 116, 157, 162)	−0.206	90	62
7	(1, 22, 53, 109, 116, 157, 162)	−0.220	121	78
8	(1, 22, 52, 71, 90, 109, 116, 159)	−0.235	177	100

[a] The seed correlation is calculated from eq 12. [b] Occurrence of the seed atom set in the 200 MC runs. [c] The number of residues covered by the local features divided by the total number of residues. Sampling time window: 2−10 ns.

output correlations (eq 11) in Figure 3b for the eight seed atoms (1, 22, 52, 71, 90, 109, 116, 159). Each seed atom is surrounded by a prominent, sequentially contiguous region of positive correlation, which forms a peak. Note that the combined output correlations closely resemble the RMSF curve (Figure 3a), demonstrating adequate coverage of the entire protein dynamics by the eight features. We can tell which seed atom is significant by the height of its peak in the output correlation plot. Besides the background noise level of small-amplitude positive correlations outside of the positive peak, there are also regions that exhibit negative correlation (not shown in Figure 3b). These anticorrelated regions are located far from the seed atoms and provide long-distance structure to the minimization criterion (eq 12) which facilitates convergence of the sparsified representations.

What is the optimal number of seed atoms in T4L? A good set of seeds should satisfy two conditions, (1) nonredundancy, that is, zero or negative output correlation between any two neighboring seed atoms, and (2) complete coverage of the protein. Besides $n = 8$ (Table 1), we found that $n = 15$ satisfies the two conditions as well. The 15 local dynamic domains (not

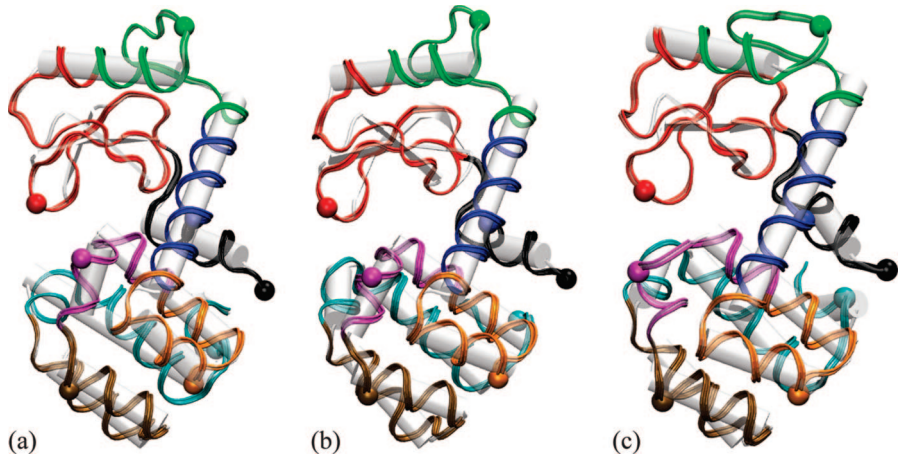

Figure 4. Locations and dynamics of the eight local features ($n = 8$) in T4L during the MD simulation. (a) Initial structure of the simulation ($t = 0$ ns), (b) $t = 4$ ns, and (c) $t = 8.25$ ns. The local dynamic domains are colored as those in Figure 3. Seed atoms are shown as spheres, and the protein in is white cartoon representation. Molecular graphics renderings were created with VMD.[36]

shown) in the case $n = 15$ are smaller than those of $n = 8$. We recommend adopting the minimum number of n that satisfies the two conditions as the optimal number of seed atoms, that is, eight local features are sufficient for T4L.

In Figure 4, the eight dynamic domains for $n = 8$ are visualized in the initial structure and in two selected snapshots of the simulation. The two terminal dynamic domains, C_α-1 and C_α-159, reflect the well-known flexibility of the open-ended chains. T4L has two major structural domains which are connected by a long α-helix (Figure 4). Both experimental and theoretical studies reveal that T4L exhibits prominent open−close and twist motions between these two structural domains.[29−32] On the basis of RMSF values and output correlation peaks, the three nonterminal domains, C_α-71 (the linking α-helix), C_α-90, and C_α-116 are less significant (termed minor) compared to the other three (termed major) nonterminal dynamic domains, C_α-22, C_α-52, and C_α-109. The two major dynamic domains, C_α-22 and C_α-109, include the cross-domain active site of T4L, which is the binding pocket with substrates. There are two hinge bending regions in the protein, which are located at both sides of the long α-helix (C_α-71). The major dynamic domain C_α-52 includes the hinge bending region near the N-terminus, and the minor dynamic domain C_α-90 comprises the hinge bending region near the C-terminus. C_α-52 is biologically more significant than C_α-90 because the known twist motion of the structural N-terminal domain relative to the structural C-terminal domain[29−32] originates near C_α-52.

The seeds found by the earlier SPA-k algorithms were interchangeable, and reordering helped ensure that the SPA-k results were robust across different portions of the MD trajectory.[15] In the case of CGLC, however, the seeds were no longer interchangeable but ordered by the linear sequence of the polypeptide chain. Therefore, it was necessary to investigate their robustness across different MD time windows. We divided the T4L trajectory into two windows (WIN I: 2−6 ns and WIN II: 6−10 ns) and assigned seeds by the CGLC algorithm as before. In the case $n = 8$, seven seed atoms were conserved between the two time windows (Table 2). Consequently, the overlap matrix between the two sets of local features was nearly diagonal (Figure 5a), in contrast to the overlap of PCA modes extracted from the two windows (Figure 5b).

An undersampling of conformations due to MD time limitations affects all PCA modes globally, rendering them unreliable. In contrast, the undersampling manifests itself locally in the LFA/CGLC overlap matrix, as can be seen in the case of seed

TABLE 2: The $n = 8$ Seed Atoms Extracted from Two Different Time Windows and from the Combined T4L Trajectory Using the CGLC Algorithm

WIN-I[a]	WIN-II[b]	overlap[c]	WIN I+II[d]
12	1	0.327	1
22	22	0.382	22
52	52	0.380	52
72	69	0.350	71
92	90	0.574	90
(162)	109	(−0.027)	109
127	116	0.110	116
159	158	0.437	159

[a] The time window from 2 to 6 ns. [b] The time window from 6 to 10 ns. [c] The overlap values between the first two columns were calculated using eq 18 in ref 15; these correspond to the diagonal elements in Figure 5a. [d] The full trajectory from 2 to 10 ns. Seeds from WIN I and WIN II were aligned to corresponding seeds from WIN I+II. One mismatch is indicated by (•••); see text.

number 6 corresponding to C_α-159 (Figure 5a). Clearly, the overlap analysis could be used as a tool to assess the convergence of local regions in the structure. Most parts of the structure appear well-sampled, and LFA seeds are conserved. The eight seed atoms from WIN II are essentially the same as those derived from the complete trajectory, WIN I+II (Table 2). The seed atoms in WIN I are also conserved, except that C_α-109 in WIN II is not found in WIN I (Table 2). This is due to large movements of C_α-109 in WIN II (Figure 4c) which are absent in WIN I (Figure 4a,b). The adjacent seed number 7 (Figure 5a), corresponding to C_α-116 (Table 2), is also affected by this undersampling, albeit to a lesser degree, exhibiting a relatively low overlap of 0.110 with the corresponding seed C_α-127 in WIN I (Table 2). Mismatches in the seed sequence and deviations in the diagonal structure of the overlap matrix thus reveal differences in the conformational sampling within the MD time windows.

Amadei et al. observed that the subspace spanned by n PCA modes converged better between different time windows as n increased.[33] More PCA modes are taken into account for LFA when $n = 15$. By performing again the overlap analysis for $n = 15$, we searched for an effect of higher n on the robustness of LFA features. Overall, the case of $n = 15$ (Table 3) is similarly well converged (13 of 15 seeds are invariant). However, the undersampling near C_α-109 and C_α-116 (observed in $n = 8$) is also apparent at $n = 15$; C_α-109 exhibits a relatively low overlap of 0.186 with the corresponding C_α-105 in WIN I,

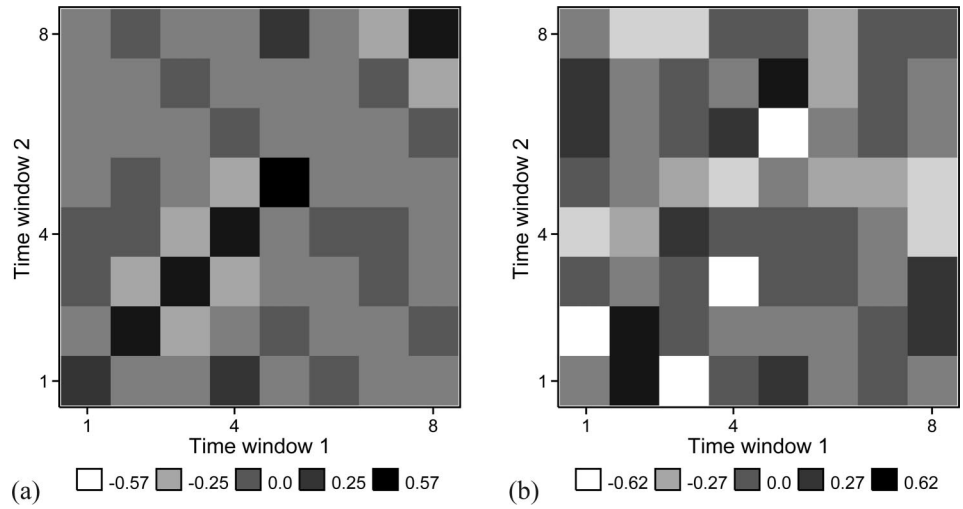

Figure 5. Overlaps between the two time windows of T4L. (a) Overlaps between the local features (represented by seed atoms) computed from eq 18 in ref 15. Eight seed atoms were determined for each time window in the order of Table 2. (b) Overlaps between the PCA modes computed from eq 17 in ref 15. The first eight PCA modes in each window were computed and sorted by descending eigenvalues.

TABLE 3: The $n = 15$ Seed Atoms Extracted from Two Different Time Windows and from the Combined T4L Trajectory Using the CGLC Algorithm

WIN-I[a]	WIN-II[b]	overlap[c]	WIN I+II[d]
1	1	0.571	1
17	15	0.343	15
22	23	0.252	22
30	28	0.327	30
38	38	0.400	38
52	51	0.505	51
56	(83)	(0.164)	56
69	68	0.364	68
92	90	0.380	90
105	109	0.186	109
(83)	115	(0.188)	115
123	124	0.552	127
135	140	0.313	137
157	156	0.461	151
162	162	0.322	162

[a] The time window from 2 to 6 ns. [b] The time window from 6 to 10 ns. [c] The overlap values between the first two columns were calculated using eq 18 in ref 15. [d] The full trajectory from 2 to 10 ns. Seeds from WIN I and WIN II were aligned to corresponding seeds from WIN I+II. Two mismatches are indicated by (•••); see text.

whereas C_α-115 is absent in WIN I altogether. Instead, there is an unstable extra feature C_α-83 which is not found in the combined trajectory. A larger n thus provides a higher level of detail in the coarse model, but the assignment of local features remains unstable for unconverged features.

Myosin Motor Domain. Myosin II motor domain S1 (MYO) is a more challenging and complex system than T4L because it is almost seven times larger and consists of multiple chains. We used a time series from 0.4 to 4.0 ns of the MD trajectory to perform LFA, and the residue numbers are given according to refs 21 and 22 to compare with experimental data. In an earlier study using a shorter 1 ns trajectory, we identified 12 seed atoms by SPA-81 to describe functional dynamics of MYO.[5] The 12 dynamic domains included most of the functional parts in MYO but did not cover the whole molecule (Figure 4b in ref 5). Here, we have investigated the cases of $n = 12$ and 22 using CGLC. We carried out MC simulations as was done for T4L, except 400 different initial seed atom sets were used, and the eight T_m values were chosen from 0.01 to 0.00005 (the maximum $T_m = 0.01$ corresponds to 10% of the peak output correlation in MYO).

The MYO system in the simulation comprises 1099 residues, and it is a much more complex biomolecular machine than T4L, with multiple functional parts.[21,22] There are six segments in the simulated MYO system (the heavy chain is subdivided into four segments by missing loops, plus the two light chains). The correlations between neighboring seed atoms h and $h + 1$ were computed for each segment separately and summed up in eq 12.

As mentioned in the section of the coarse-graining method, the computational complexity of optimizing n seed atoms among N atoms is given by the binomial coefficient $\binom{N}{n}$. For example, in the case of T4L when $n = 8$ and $N = 162$, this number is about 10^{13}. However, in the case of MYO, $n = 12$ and $N = 1099$ gives a number of about 10^{27} combinations, yielding a search space that is 14 orders of magnitude larger. Therefore, the MC simulations of MYO are less likely to converge than T4L, as there are many nearly degenerate local minima with minor seed correlation differences in the global search space.

To compare with SPA-81,[5] we determined 12 seeds from the new MYO trajectory using GCLC. The lowest-correlation set (67, 204, 371, 410, 501, 666, 737, 840, 859, 991, 1063, 1144) appeared 52 times in the 400 runs. Out of the 12 seeds, 3 were identical (204, 410, 737) to the earlier SPA-81, five seeds (GCLC: 67, 501, 859, 991, 1063) were very similar (SPA- 81: 56, 492, 861, 1008, 1060), and two GCLC seeds (666, 1144) were very similar to SPA-81 seeds (667, 1140) extracted from an enhanced sampling trajectory.[5] Only 2 of the 12 seeds (371, 840) differed significantly from the earlier SPA-81 results due to the absence of the artificial neighborhood exclusion in GCLC. All of the 12 CGLC seed atoms exhibit relatively high RMSF values (Figure 6a) that correspond to the flexible parts in the protein, and the 12 dynamic domains cover about 79% of the system. Eight out of 12 local features are located within the heavy chain, and the other 4 are in the light chains (2 in the RLC and 2 in the ELC).

We found $n = 22$ was the minimum number of seeds required to cover the entire MYO system without redundancy. The 22 seed atoms are (4, 67, 204, 333, 370, 410, 452, 571, 576, 625, 681, 736, 821, 860, 901, 923, 951, 1008, 1016, 1063, 1084, 1142), a set that appeared only once in the 400 runs. There are 13 local features in the heavy chain, 5 in the RLC, and 4 in the

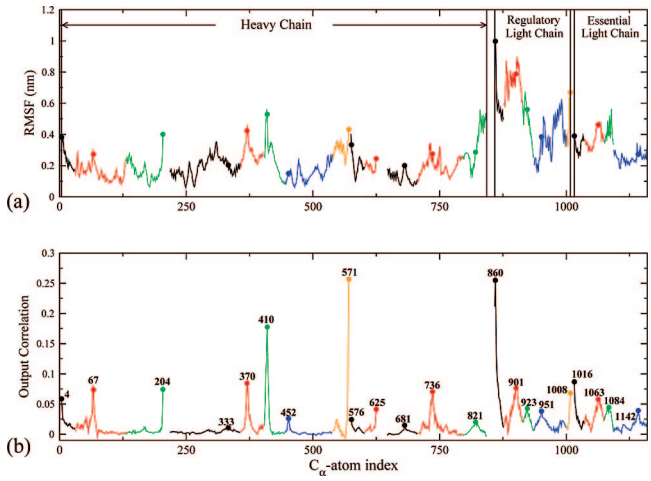

Figure 6. (a) RMSF of C_α atoms calculated from the MD simulation of MYO. The RMSF curve is divided into the 22 local features. C_α-4 (4−30) black; C_α-67 (31−131) red; C_α-204 (132−204) green; C_α-333 (218−356) black; C_α-370 (357−403) red; C_α-410 (404−444) green; C_α-452 (445−538) blue; C_α-571 (539−571) orange; C_α-576 (575−602) black; C_α-625 (603−626) red; C_α-681 (647−706) black; C_α-736 (707−795) red; C_α-821 (796−843) green; C_α-860 (859−876) black; C_α-901 (877−913) red; C_α-923 (914−935) green; C_α-951 (936−1000) blue; C_α-1008 (1001−1008) orange; C_α-1016 (1016−1035) black; C_α-1063 (1036−1074) red; C_α-1084 (1075−1093) green; and C_α-1142 (1094−1160) blue. The heavy chain, RLC, and ELC are indicated. There are four segments in the heavy chain due to missing residues. (b) Output correlations (eq 11) between the seed atoms and MYO as a function of residue number in the case of $n = 22$. Only the correlations between one seed atom and the atoms in its corresponding dynamic domain are plotted, which are colored according to the RMSF curve. The 22 seed atoms are indicated by circles. The residue numbers are from refs 21 and 22.

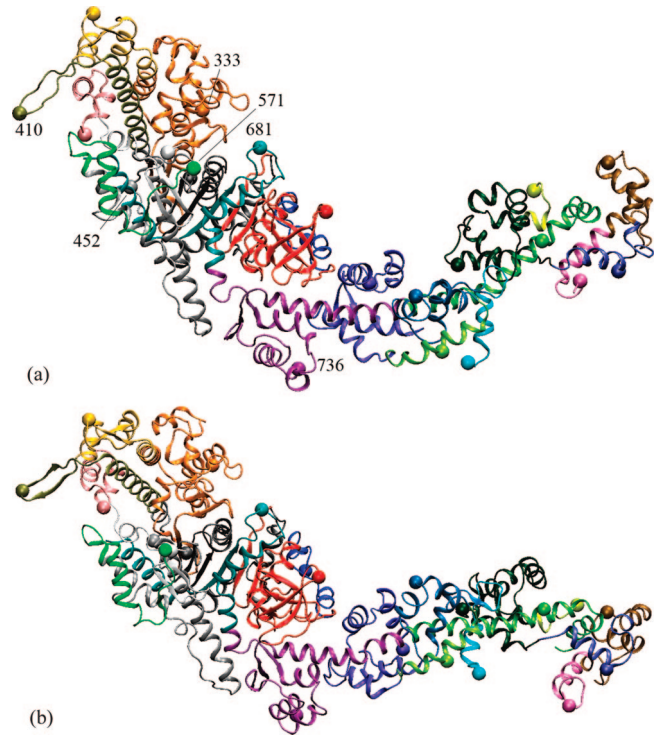

Figure 7. Locations and dynamics of the 22 local features ($n = 22$) in MYO during the MD simulation. (a) Initial structure of the simulation ($t = 0$ ns) and (b) the final structure of the simulation ($t = 4$ ns). Seed atoms are shown as spheres, and local dynamic domains in cartoon representation are colored accordingly. Each domain contains a localized and contiguous set of atoms that have positive output correlations (eq 11) with the corresponding seed atom (Figure 6). Specific seed atoms discussed in the main text are labeled in (a). C_α-4 blue; C_α-67 red; C_α-204 gray; C_α-333 orange; C_α-370 yellow; C_α-410 tan; C_α-452 silver; C_α-571 green; C_α-576 white; C_α-625 pink; C_α-681 cyan; C_α-736 purple; C_α-821 lime; C_α-860 mauve; C_α-901 ochre; C_α-923 iceblue; C_α-951 black; C_α-1008 yellow2; C_α-1016 cyan2; C_α-1063 blue2; C_α-1084 yellow3; and C_α-1142 violet. Molecular graphics renderings were created with VMD.[36]

ELC. The RMSF values of the seed atoms are presented in Figure 6a, and the corresponding dynamic domains associated with the seed atoms are identified in Figure 6b, which demonstrates that the dynamic domains are allocated predominantly in the most flexible regions of the system.

In Figure 7, we visualize the local features in the initial (a) and final (b) structure of the MD simulation. The myosin motor protein is a molecular machine with many allosterically coupled functional units,[34] and the local dynamics domains that we identified can be well related to these functionally well-known parts. For example, Ala333 and Lys681 are near the entry for the active site where ATP is hydrolyzed and chemical energy is released and turned into mechanical motion. Asn410 is located in the so-called "cardiomyopathy loop", whose disruption by missense mutations is implicated in the familial hypertrophic cardiomyopathy.[35] The local domain represented by seeds Lys452 and Ala571 are close to the Actin-binding domain, and the local feature defined by Gln736 is the converter domain of MYO at the start of the lever arm. The local features in the two light chains, RLC and ELC, cradle myosin's lever arm.

We divided again the trajectory into two time windows (WIN-I: 0.4−2.2 ns and WIN-II: 2.2−4.0 ns) and identified 22 seed atoms in both windows by CGLC. The results in Table 4 indicate that the seed atoms are quite similar between the different time windows (of the 22 seeds, 15 are conserved), although the short simulation is almost certainly not converged (4 ns is a very short simulation relative to the millisecond time scale of the myosin power stroke). The seven mismatched seeds in Table 4 are indicative of conformational undersampling in the short MD trajectory. For example, three of the light chain seeds are not conserved. A significant bending motion of the lever arm and the light chains is observed in the simulation (Figure 7). Despite their small size, a relatively large number of seeds (nine) is assigned to the RLC and ELC because the light chains fluctuate significantly in the MD simulation (Figure 6a). Small variations of the number of light chain seeds between the time windows are therefore expected due to the unconverged lever arm motion.

The diagonal structure of the overlap matrix between the local features (Figure 8a) shows considerable improvement compared to the PCA overlap matrix (Figure 8b). As before in the case of T4L, the matrix shows that most features are conserved. The nondiagonal or weakly diagonal elements are due to conformational undersampling. Specifically, seeds 4 and 7 in the matrix, corresponding to Ala333 and Lys452, are not diagonal. This region is located in the 50 K domain of myosin that contains the ATP- and Actin-binding sites (Figure 7), suggesting that 4 ns was too short of a duration to sample all relevant motions in these two locations.

In summary, the above results suggest that the coarse-graining by CGLC provides robust local features also for larger sized systems such as MYO, whose conformational variability is almost certainly undersampled in a MD simulation on the nanosecond time scale. The coarse-grained convergence analysis allows one to visualize the location of both well- and undersampled regions, a major advantage over nonlocal PCA convergence analysis.[14]

TABLE 4: The $n = 22$ Seed Atoms Extracted from Two Different Time Windows and the Combined MYO Trajectory Using The CGLC Algorithm

WIN-I[a]	WIN-II[b]	overlap[c]	WIN I+II[d]
20	5	0.117	4
66	66	0.191	67
204	(560)	(0.203)	204
(57)	(256)	(−0.059)	333
(307)	(991)	(0.101)	370
409	410	0.220	410
(74)	(729)	(−0.077)	452
571	571	0.116	571
575	575	0.193	576
615	625	0.186	625
663	657	0.250	681
737	718	0.249	736
840	843	0.222	821
859	859	0.228	860
900	901	0.264	901
923	919	0.214	923
951	(808)	(0.309)	951
1008	(141)	(0.160)	1008
1018	1016	0.244	1016
1063	1065	0.344	1063
1084	(370)	(0.122)	1084
1142	1109	0.207	1142

[a] The time window from 0.4 to 2.2 ns. [b] The time window from 2.2 to 4 ns. [c] The overlap values between the first two columns were calculated using eq 18 in ref 15; these correspond to the diagonal elements in Figure 8a. [d] The full trajectory from 0.4 to 4 ns. Seeds from WIN I and WIN II were aligned to corresponding seeds from WIN I+II. Seven mismatches are indicated by (•••), see text.

Conclusions

In this article, we present a novel coarse-graining algorithm, CGLC, for the assignment of local dynamic features from MD trajectories. Instead of a sequential optimization of the seed atoms in SPA-k ($k \geq 0$) subject to k-neighbor exclusion, we perform a stochastic minimization of the linear chain seed correlation (eq 12). Applications of CGLC to two biological systems, T4L and MYO, demonstrate that CGLC overcomes the major earlier limitations, redundancy of SPA-0 and artificial neighborhood exclusion of SPA-k ($k > 0$), providing a purely statistical solution to the coarse-graining problem.

In the T4L case, eight seed atoms (1, 22, 52, 71, 90, 109, 116, 159) computed by CGLC can be well related to functional domain motions of the protein (Figure 4). For such small systems, individual MC minimizations of the seed correlation converge rapidly to the optimal solution.

In practical applications of LFA, the question arises how many local features are needed to account for the dynamics. It is hard to estimate the ideal number n using the earlier algorithms SPA-k ($k \geq 0$). For example, in ref 15, we computed up to 15 seed atoms for T4L. Using CGLC, we are now able to define a minimum number required to cover the entire system without redundancy. For the small T4L system, eight local features are sufficient to describe its dynamics. In general, we expect this ideal number of features to depend on the system size and length of the MD simulation.

When dealing with a large size system such as MYO, there is a problem of convergence. Several factors conspire to complicate the search for a statistically reproducible representation. First, the complexity of picking n seeds out of N atoms increases dramatically both with the system size and the number of seeds, which is larger for multidomain biomolecular machines. Second, the output correlations become more localized[15] as n increases. For example, the MYO output correlations in Figure 6b exhibit significantly sharper peaks compared to those of the T4L correlations in Figure 3b. Therefore, the seed correlation landscape (eq 12) is frustrated, and individual MC runs might get trapped in many local minima. The CGLC algorithm does not guarantee that the MC search finds the global optimum. However, the observed local features of MYO corresponded to well-known functional units in this molecular machine and were largely conserved across MD time windows.

The well-known "MD sampling problem"[13] is apparent also in our model, in the sense that some localized features differ among MD sampling windows. The local convergence can be quickly assessed by means of the output overlap matrix (Figures 5 and 8). The new coarse-graining offers a major advantage; unlike in PCA, where modes are related by orthogonality and easily perturbed, the converged LFA features are invariant across multiple windows of the trajectory, dividing the protein into converged regions and a smaller number of localized, undersampled regions. The proposed coarse-graining thus provides a localized measure of MD sampling efficiency.

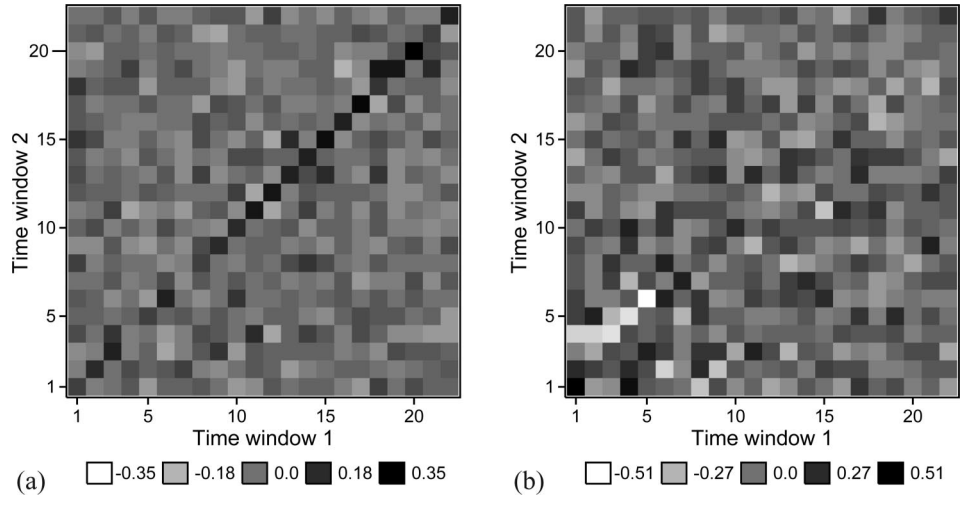

Figure 8. Overlaps between the two time windows of MYO. (a) Overlaps between the local features (represented by seed atoms) computed from eq 18 in ref 15. Twenty-two seed atoms were determined for each time window in the order of Table 4. (b) Overlaps between the PCA modes computed from eq 17 in ref 15. The first 22 PCA modes in each window were computed and sorted by descending eigenvalues.

Summing up, the goal of this article was to develop a robust coarse-graining technique for routine trajectory analysis and to provide a practical road map for the application of LFA. We expect a user would take the following steps when applying LFA:

1. Perform PCA and order the principal components by decreasing eigenvalues.

2. Decide on an initial n (usually a small number at the beginning) of local features and then compute the $P = K^0$ matrix (eq 3) from the first n PCA modes.

3. Generate a number of random initial seeds, estimate the range of T_m values (the maximum T_m can be about 10% of the peak output correlation, and then decrease it to obtain other T_m values), and perform individual MC minimizations of eq 12 for each set until converging to its lowest seed correlation set. For a complex protein assembly like MYO, every chain should be treated separately, which means the correlation between the seed atoms h and $h + 1$ is not minimized if they are not in the same chain.

4. Perform steepest descent on each set after MC. For multichain systems, the seed atoms are not allowed to cross chain boundaries.

5. Pick out the lowest seed correlation set from all of the sets and assign a dynamic domain for each seed atom (contiguous region of positive output correlation). If the domains cover the whole molecule without redundancy, this set would be the final result of CGLC; otherwise, increase n and repeat from step 2 to get more coverage.

LFA is a novel trajectory analysis technique that opens the door to further studies of "essential" dynamic features in proteins. We expect that the implementation details given in this article facilitate future applications by other groups.

Acknowledgment. This work was supported by grants from NIH (R01GM62968), Human Frontier Science Program (RGP0026/2003), and the Alfred P. Sloan Foundation (BR-4297).

References and Notes

(1) Brooks, C. L., III; Karplus, M.; Pettitt, B. M. Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics; *Advances in Chemical Physics*; John Wiley & Sons: New York, 1988; Vol. LXXI.

(2) Karplus, M. Molecular Dynamics: Applications to Proteins. In *Modelling of Molecular Structures and Properties*; Proceedings of the 44th International Meeting of Physical Chemistry, Nancy, France, September 11−15, 1989; Rivail, J.-L., Ed.; Elsevier Science Publishers: Amsterdam, The Netherlands, 1990, Vol. 71.

(3) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(4) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589–1615.

(5) Wriggers, W.; Zhang, Z.; Shah, M.; Sorensen, D. C. *Mol. Simul.* **2006**, *32*, 803–815.

(6) Lange, O. F.; Grubmüller, H. *J. Phys. Chem. B* **2006**, *110*, 22842–22852.

(7) Amadei, A.; Linnsen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412–425.

(8) van Aalten, M. F.; Amadei, A.; Linssen, A. B. M.; Eijsink, V. G. H.; Vriend, G.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1995**, *22*, 45–54.

(9) Kitao, A.; Go, N. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–169.

(10) Berendsen, H. J. C.; Hayward, S. *Curr. Opin. Struct. Biol.* **2000**, *10*, 165–169.

(11) Hernández, G.; Jenney, F. E., Jr.; Adams, M. W. W.; LeMaster, D. M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 3166–3170.

(12) Falke, J. J. *Science* **2002**, *295*, 1480–1481.

(13) Clarage, J. B.; Romo, T.; Andrews, B. K.; Pettitt, B. M.; Phillips, G. N. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3288–3292.

(14) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *100*, 2567–2572.

(15) Zhang, Z.; Wriggers, W. *Proteins: Struct., Funct., Bioinf.* **2006**, *64*, 391–403.

(16) Lange, O. F.; Grubmüller, H. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1294–1312.

(17) Penev, P. S.; Atick, J. J. *Network-Computation in Neural Systems* **1996**, *7*, 477–500.

(18) Metropolis, N.; Rosenbluth, M.; Rosenbluth, A.; Teller, A.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(19) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; BIOMOS b. v.: Zürich, Switzerland; Groningen, The Netherlands, 1996.

(20) Scott, W. R. P.; Hnenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.

(21) Holmes, K. C.; Angert, I.; Kull, F. J.; Jahn, W.; Schröder, R. R. *Nature* **2003**, *425*, 423–427.

(22) Rayment, I.; Rypniewski, W. R.; Schmidt-Bäse, K.; Smith, R.; Tomchick, D. R.; Benning, M. M.; Winkelmann, D. A.; Wesenberg, G.; Holden, H. M. *Science* **1993**, *261*, 50–58.

(23) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981.

(24) Verlet, L. *Phys. Rev.* **1967**, *159*, 98–103.

(25) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(26) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(27) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

(28) Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1994**, *100*, 3169–3174.

(29) Faber, H. R.; Matthews, B. W. *Nature* **1990**, *348*, 263–266.

(30) Zhang, X. J.; Wozniak, J. A.; Matthews, B. W. *J. Mol. Biol.* **1995**, *250*, 527–552.

(31) de Groot, B. L.; Hayward, S.; van Aalten, D. M. F.; Amadei, A.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1998**, *31*, 116–127.

(32) Zhang, Z.; Shi, Y.; Liu, H. *Biophys. J.* **2003**, *84*, 3583–3593.

(33) Amadei, A.; Ceruso, M. A.; Di Nola, A. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 419–424.

(34) Geeves, M. A.; Holmes, K. C. *Annu. Rev. Biochem.* **1999**, *68*, 687–728.

(35) Fananapazir, L.; Dalakas, M. C.; Cyran, F.; Cohn, G.; Epstein, N. D. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3993–3997.

(36) Humphrey, W. F.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

JP806291P