ELSEVIER

Physica D 107 (1997) 225–239

PHYSICA D

Multi-basin dynamics of a protein in a crystal environment

Angel E. García [a,*], Raphael Blumenfeld [b,c], Gerhard Hummer [a,b], James A. Krumhansl [d]

[a] *Theoretical Biology and Biophysics Group, T-10, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

[b] *Theoretical Division and Center for Nonlinear Studies (CNLS), Los Alamos National Laboratory, Los Alamos, NM, 87545, USA*

[c] *Cambridge Hydrodynamics, Princeton, NJ, USA*

[d] *Laboratory of Atomic and Solid State Physics, Department of Physics, Cornell University, Ithaca, NY 14854, USA*

Abstract

The dynamics of the small protein crambin is studied in the crystal environment by means of a 5.1 nanoseconds molecular dynamics (MD) simulation. The resulting trajectory is analyzed in terms of a small set of nonlinear dynamical modes that best describe the molecule's fluctuations. These modes are nonlinear in the sense that they describe a trajectory exhibiting multiple transitions among local minima at various timescales. Nonlinear modes are responsible for most of the protein atomic fluctuations. An ultrametric hierarchy of sampled local minima describes the protein trajectory when structures are classified in terms of their interconfigurational mean squared distance. Transitions among minima involve small changes in the relative atomic positions of many atoms in the protein. The character of the MD trajectory fits within the framework of rugged energy landscape dynamics. This MD simulation clarifies the unique statistical features of the barriers between minima in the energy-like configurational landscape. Longer timescale dynamics seem to sample transitions between minima separated by relatively higher barriers. The MD trajectory of the system in configurational space can be described in terms of diffusion of a particle in real space with a waiting time distribution due to partial trapping in shallow minima. A description of the dynamics in terms of an open Newtonian system (the protein) coupled to a stochastic system (the solvent and fast quasiharmonic modes of the protein) reveals that the system loses memory of its configurational space within a few picoseconds. The diffusion of the protein in configurational space is anomalous in the sense that the mean square displacement increases sublinearly with time, i.e., as a power law with an exponent that is smaller than unity.

Keywords: Protein dynamics; Nonlinear dynamics; Ultrametric hierarchy; Molecular dynamics; Anomalous diffusion; Lévy flights

1. Introduction

The dynamics of proteins are closely related to their biological function. Proteins exhibit motions on a very wide range of timescales from picoseconds (ps) [1–3] to seconds [4] (as studied by hydrogen exchange experiments). Experimental studies on myoglobin suggest the existence of a hierarchy of motions occurring at various timescales

* Corresponding author. Tel.: (505) 665-5341; fax: (505) 665-3493; e-mail: angel@t10.lanl.gov.

PII S0167-2789(97)00090-0

and resulting from an ensemble of nearly degenerate states separated by a distribution of enthalpic energy barriers [5–8]. On longer timescales (milliseconds to seconds) many proteins seem to follow an energy-like funnel to the folded state [9,10]. Fast folding (submillisecond regime) proteins have also been observed [11–14]. Studies of the protein folding kinetics and stability analyses based on 3D lattice and off-lattice models of polymers with binary interactions (i.e., polar and hydrophobic) that mimic proteins [15–21] also agree with this picture. Onuchic, Wolynes and collaborators [16,22–24] proposed an energy landscape theory of folding with the idea that folding kinetics is best regarded as a progressive organization of an ensemble of partially folded structures, the *folding funnel*, rather than a serial progression between intermediates. Recent experiments [9,25–27] show that varying one crucial amino acid can eliminate a late stage folding bottleneck and allow fast folding to occur, in perfect agreement with the folding-funnel theory.

Krumhansl [28] related the dynamics of proteins and DNA to that of low-dimensional materials showing structural phase transitions. In support of these observations García [29] showed that the fluctuations of a protein in solution are best described in terms of large-amplitude nonlinear motions. Molecular dynamics (MD) simulations, ligand recombination [30], pressure relaxation [7,8], time resolved X-ray crystallography [31] and vibrational echo studies [1–3] give information about the lower end of the funnel or energy landscape that is sampled in the folded state. Computational evidence (MD and Monte Carlo (MC) simulations) for the existence of these substates have been previously discussed in the literature [29,32–34].

Here we will present evidence showing the presence of multi-basin, nonlinear motion in proteins in the picoseconds (ps) and nanoseconds (ns) timescales. A method for extracting modes that best represent the fluctuations in the system will be described. This method consists of a generalized least-square fitting of sampled configurations in Cartesian space to one-, two-, or three-dimensional subspaces that best describe the atomic fluctuations [29]. We will show that the MD trajectory of the protein is clustered around few local minima (basins of attraction), and that many transitions among local minima occur within the 5 ns trajectory. These transitions involve small changes in the relative atomic positions of many atoms in the protein. The trajectory of the protein in configurational space can be described within a framework of a particle diffusing in real space and getting partially trapped in shallow minima. A description of the dynamics in terms of an open Newtonian system (the protein) coupled to a stochastic system (the solvent and fast quasiharmonic modes of the protein) reveals that the system loses memory of its configuration within a few ps. The diffusion of the protein state in configurational space is anomalous in the sense that the mean square displacement, $\langle R^2 \rangle$, is not proportional to time, as in traditional Brownian diffusion, but is substantially suppressed with $\langle R^2 \rangle \sim t^{2H_D}$, where $2H_D < 1$.

2. Description of the system

We study the dynamics of a small hydrophobic protein, crambin, in its crystal environment during a 5.1 ns, 300 K MD simulation. Crambin is a 46 amino acid protein that contains most of the structural elements that are characteristic of larger proteins. Fig. 1 shows a schematic representation [35] of the three-dimensional structure of crambin. Starting from the N-terminus and moving along the protein chain we find a β-strand (S_1, amino acids 1–4), a loop (amino acids 5–6), a helix (H_1, amino acids 7–19), a loop (amino acids 20–22), another helix (H_2, amino acids 23–30), another β-strand that makes hydrogen bonds with the first β strand to form a β sheet (S_2, amino acids 32–35), and a turn (amino acids 41–44). Three disulfide bonds are formed by Cys(3)–Cys(40), Cys(4)–Cys(32), and Cys(16)–Cys(26). Because of these disulfide bonds the connectivity of the amino acid chain cannot be described by a quasi-one-dimensional chain.

The initial conformation of the protein was obtained from the crystallographic coordinates reported by Hendrickson and Teeter [36]. In the crystal, crambin adopts a $P2_1$ space group symmetry, contains two molecules

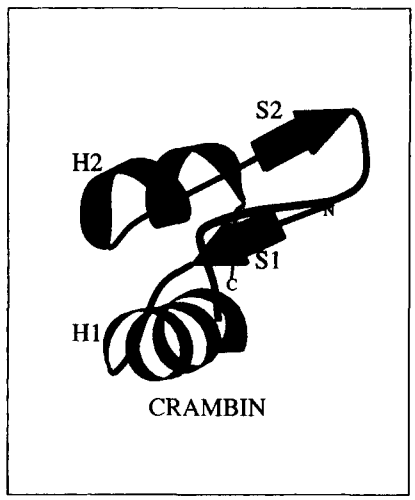

Fig. 1. Ribbon representation of the secondary and tertiary structure of crambin. The letters N and C label the amino and carboxy termini, respectively. The labels H1 and H2 show two helical regions. The labes S1 and S2 show the two β structures. This figure was generated using the program MOLSCRIPT [35].

per unit cell, and the unit cell dimensions are $a = 40.96$, $b = 18.65$, $c = 22.52$ Å, with $\beta = 90.77°$. We simulated the dynamics of two molecules in the unit cell without imposing the $P2_1$ symmetry, except on the initial configuration. We used periodic boundary conditions to simulate an infinite crystal where all the unit cells are perfectly correlated. We added 182 water molecules to the system according to crystal density estimates by Teeter [37] and MC simulations by Jorgensen et al. [38]. Interatomic energy interactions were modeled with the all-atom force field of Cornell et al. [39]. This force field approximates intra- and intermolecular interactions through classical potentials. For water we used the TIP3P model [40]. The simulated system thus contains 1830 atoms in the unit cell (182 water molecules and two crambin molecules with 642 atoms each). We calculated long-ranged electrostatic energies using the particle-mesh-Ewald-summation (PME) method [41], implemented in Amber [42]. $64 \times 32 \times 32$ grid points were used for the PME. A third-order spline interpolation was used and the real space tolerance factor was set to 1.1×10^{-6}. Configurational averages are calculated from configurations saved at a rate of 10 per ps during the first 1 ns of the simulation and 4 per ps during the last 4 ns. We simulated the system for 5.1 ns at constant temperature [43]. The first 100 ps of the simulation are not included in any averaging to avoid artifact transitory effects. Descriptions of the dynamics of crambin in aqueous (non-crystalline) solution have been reported before [29,44,45].

3. Results and discussion

3.1. Time evolution of the distance between configurations

The inter-dependence of local structural-variables describing collective, delocalized excitations is not trivial and a description of the dynamics of a protein in terms of non-structural variables is desired. To do so we use the N-particle root-mean-square (rms) distance [46], $d(t, t')$, between evolving protein configurations to represent the fluctuations of the system. Given two configurations of the molecule, $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t')$, with centroids at the origin, the *mean square* distance is defined as the minimum of the residual, $d^2(t, t') = (1/N) \sum^N (\boldsymbol{x}(t) - \boldsymbol{x}'(t'))^2$, where

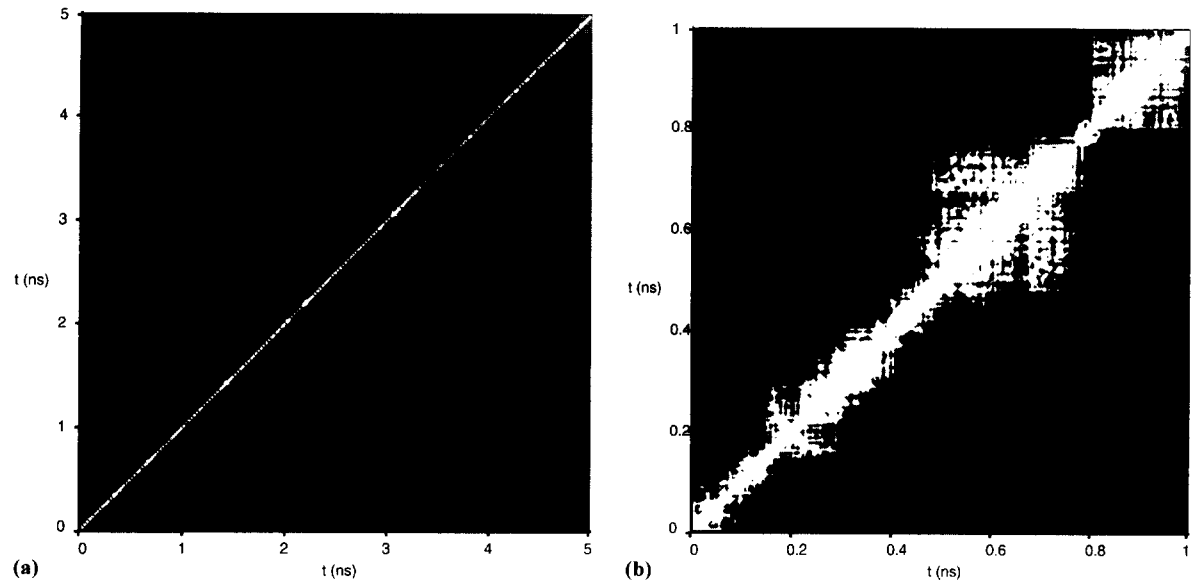

Fig. 2. (a) Contour plot of the rms distance between pairs of conformations adopted by the protein every 25 ps along the 5 ns MD trajectory. Regions surrounded by the contours are shaded from white ($d \approx 0.5$ Å) to black ($d \geq 1.1$ Å). The largest rms distance is 1.15 Å. (b) Contour plot (at 0.1 Å intervals) of the rms distance between pairs of conformations adopted by the protein every 5 ps along the first 1 ns MD trajectory.

$x'_{i,n}(t') = \sum_j R_{ij} x_{j,n}$, N is the number of atoms in the molecule, R_{ij} are the elements of an orthogonal rotation matrix with determinant $+1$, and $x_{i,n}$ is the ith component of the nth atom in the molecule. A large rms distance between configurations at short time difference $t - t'$ is indicative of fast configurational changes.

The distance matrix $d(t, t')$ between pairs of conformations at t and t', sampled every 25 ps, during the last 5 ns of simulation, is shown in Fig. 2(a). A darker gray shading implies a larger rms distance between pairs of configurations and vice versa. The rms distance smoothly increases from 0 to values near 0.5 Å in a short time (50.0 ps), and reaches a distance of near 1.0 Å after 1–2 ns. Oscillations between larger (1–1.1 Å) and smaller (0.75 Å) rms distances occur also at intervals of 200 ps. Normal mode analyses of the protein dynamics shows the lowest harmonic frequency modes to have periods of the order of a few ps [47,48]. Therefore the motions responsible for the oscillations seen here must originate from non harmonic collective long-range behavior. Fig. 2(b) shows a similar distance matrix, but this time over configurations sampled every 5 ps during the first 500 ps. Notice that a portion of the plot that appears featureless in Fig. 2(a) exhibits similar oscillations when viewed on this finer scale. Similar features are also seen when configurations are sampled over even shorter (1 ps) timescales [29]. These features suggest a *hierarchy* of motions that occur over various timescale decades.

3.1.1. Tree analysis

The information contained in $d(t, t')$ is sufficient to construct a hierarchical representation of configurations adopted by the system during the simulation and saved at fixed time intervals. The branching of such a tree will be indicative of the proximity of one configuration to another. To build the hierarchy we use the following clustering algorithm [49]:

1. Start with M configurations and a distance matrix, $d(t, t')$, containing the rms distance among all pairs of configurations. At this stage, each configuration belongs to a separate cluster.

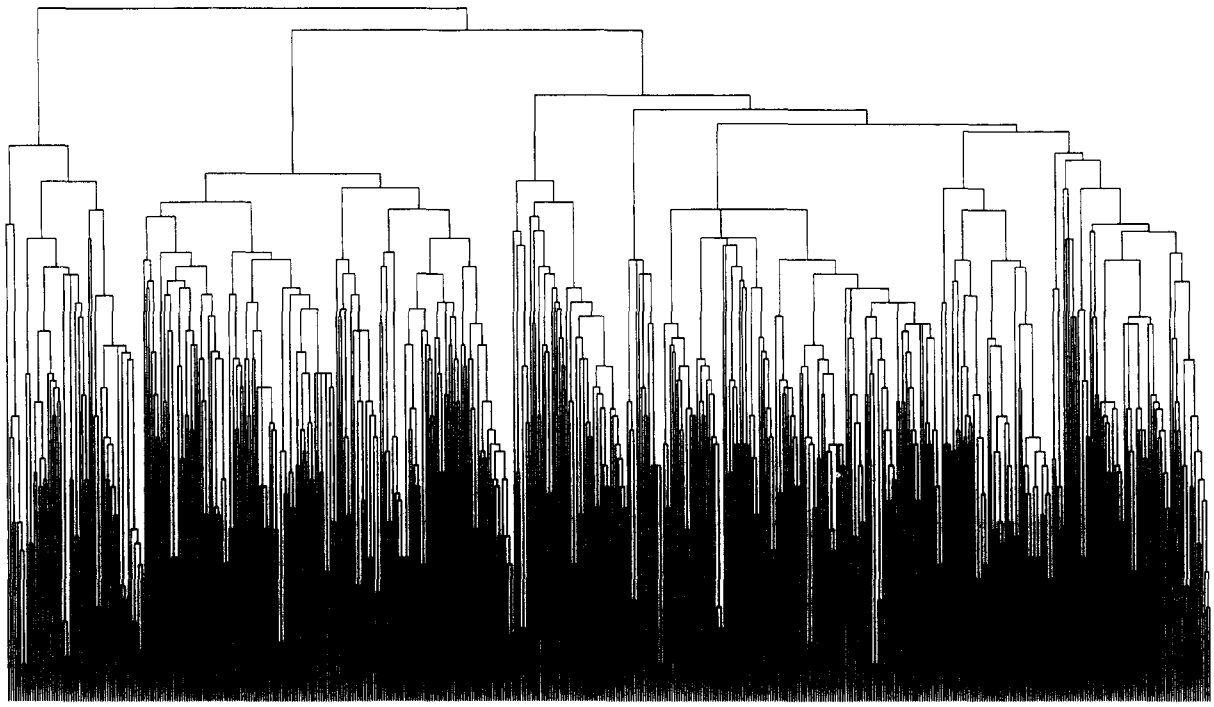

Fig. 3. Hierarchical tree representation of structures (occuring at 5 ps intervals) into different clusters. The distance among clusters is proportional to the distance between nodes on the horizontal axes.

2. Join into one cluster, l, the two distinct ($d \neq 0$) clusters (i, j) for which $d(i, j)$ is the smallest. Now we have $M - 1$ clusters. To construct the new distance matrix we take $d^{(M-1)\times(M-1)}(l, k) = \min[d(i, k), d(j, k)]$. This step is repeated $M - 1$ times, until only one cluster remains. The resulting hierarchy is graphically represented by joining each pair of newly clustered configurations by a line of length proportional to the distance between the two clustered structures. The resulting hierarchy can be indexed by the distance between clusters. Any indexed hierachy can be proven to be ultrametric if we choose the distance between two clusters in the hierarchy to be the minimum of the distances among members of each cluster [49].

Fig. 3 shows the hierarchical tree obtained by this algorithm. Branches emerging from nodes represent a family of structures that are closely related, i.e., they represent configurations in nearby local minima, while members of different families are configurations in far-away minima. This hierarchical tree conforms to the idea of a hierarchically rugged energy-like landscape proposed by Frauenfelder et al. [5], where transitions between structures in nearby minima are fast, while transitions between far-away states are reached through series of multiple jumps to intermediate nearby minima. The tree presented here is just the bottom of a hierarchy representing the energy-like landscape; i.e., it covers structures with small differences in the position of a few atoms, to structures differing in the relative orientation of helices and turns. The complete hierarchy may extend all the way from fully folded structures either to structures that exhibit completely different folding or to totally unfolded structures. The hierarchy presented here has been constructed to satisfy ultrametricity, i.e., to say, the distances between clusters satisfy $d(i, j) \leq \min[d(i, k), d(j, k)]$. However, it should be noted that ultrametricity is a consequence of the choice made for the definition of the distance between clusters in the second step of the clustering algorithm and need not reflect an inherent property of the biomolecular system.

3.2. Molecule optimal dynamic coordinates (MODC)

The oscillations shown in Fig. 2 and the branching of the tree in Fig. 3 represent collective nonlinear motions [28]. To show this we define a set of directions, $\boldsymbol{m}$, in the $3N$-dimensional space of the protein that best represent (in a least-square sense) fluctuations of the protein structure. These "molecule optimal dynamic coordinates" (MODC) have been previously described [29,45,50,51]. Motions along these directions show multi-centered oscillations, rapid transitions from one center to another, and damped quasiharmonic oscillations around each center.

A generalization of this method to represent two-dimensional and three-dimensional cuts of the configurational space, as planes and volumes, that better represent the dynamics of the system has also been presented previously [44]. Similar methods have been employed by Amadei et al. [52] to describe what they called the biomolecule's essential modes, and by Gō et al. [53] who call them principal modes. These coordinates are specific to the molecule and trajectory sampled during an MD simulation. The directions $\boldsymbol{m}^{3N}$ are determined by minimizing the mean square distances of the $\{\boldsymbol{r}_i^{3N}\}$ configurations *normal* to $\boldsymbol{m}^{3N}$, such that most of the fluctuations will be along $\boldsymbol{m}^{3N}$. The derivation of this formalism is presented in Appendix A. We calculate the MODC for a system of two molecules

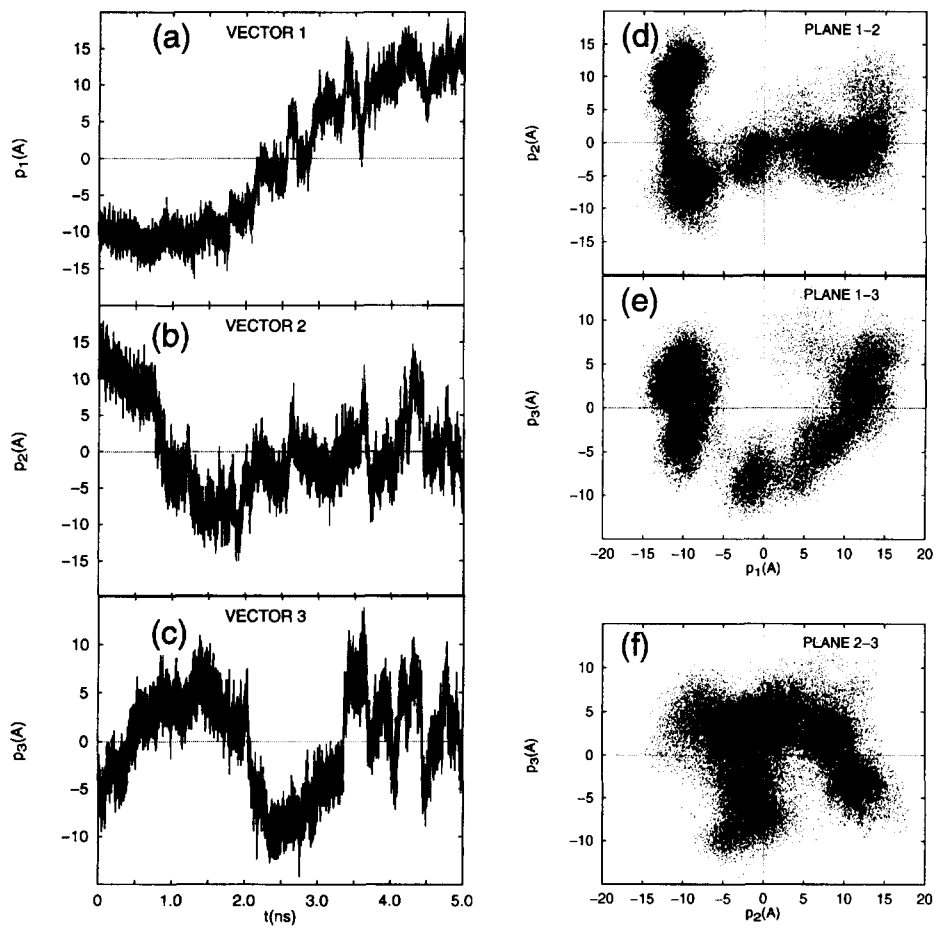

Fig. 4. Projections of the 5.0 ns MD trajectory along the three principal MODC is shown on the left panels (a–c). Projections of the trajectory on planes spanned by two MODC are shown on the right panels (d–f).

in the unit cell, thus including intermolecular motions in the characterization of the dynamics. Water molecules are not included in this analysis. The total number of atoms included in the analysis is $N = 1308$.

Fig. 4 shows the projection of the trajectory along the three principal MODC. The system shows mean square atomic fluctuations of 0.5 Å^2 during the simulation. MODC 1–3 describe 30%, 12%, and 9% of the fluctuations. The first five MODC describe 62% and the first 10 describe 74% of the total fluctuations. That is, a very small subset of MODC can represent most of the system fluctuations! This observation has far reaching implications concerning the analysis of the system, as it enables us to reduce by *many orders of magnitude* the number of degrees of freedom that are necessary for describing the system into only manageably few. Fast inter-basin transitions followed by overdamped oscillations (and possibly transitions to other local minima within each basin of attraction) can be observed along these MODC. Projections of the trajectory along the ith mode are labeled p_i. Large changes in conformation are detected at 0.8 (MODC-2), 2.0 (MODC-3), and 3.5 ns (MODC2 and 3). The right-hand side panels show the projection of the MD trajectory on planes spanned by only two MODC. The plane on top is spanned by MODC 1 and 2, and is the plane that best represents the fluctuations of the system. From this projection it is clear that the system gets trapped in four main basins near $(p_1, p_2) = (-10, 10)$, $(-10, -10)$, $(0, 0)$ and $(10, 0)$ Å. A projection of the trajectory on the second best plane, spanned by MODC 1 and 3, shows that these basins are further separated into other basins. This is better illustrated in a 3D projection of the trajectory on a volume spanned by the first three MODC shown in Fig. 5.

It must be clarified that the MODC analysis yields different eigenvalues and eigenvectors when different segments of the trajectory are analyzed. Therefore, a short simulation cannot give an estimate of the most relevant motions of the system in a larger simulation. However, a small subset of MODC (~ 10) describes most of the fluctuations of the system during a longer simulation, although it gives different eigenvalues (i.e., amplitude of the fluctuations) [44]. One can expect that the volume covered by a subset of eigenvectors from different time segments is essentially the same although viewed from a different perspective. For instance, if we consider the first 2 ns of the trajectory shown in Fig. 4, we will find that MODC 2 and 3 will be the dominant modes, while MODC 1 will oscillate around an amplitude of -10 Å. The average conformation will also change. The MODC analysis provides a quantitative tool to partition the multi-dimensional configurational space into smaller subspaces that best describe the fluctuations of the system and the topology of the energy surface sampled during a trajectory.

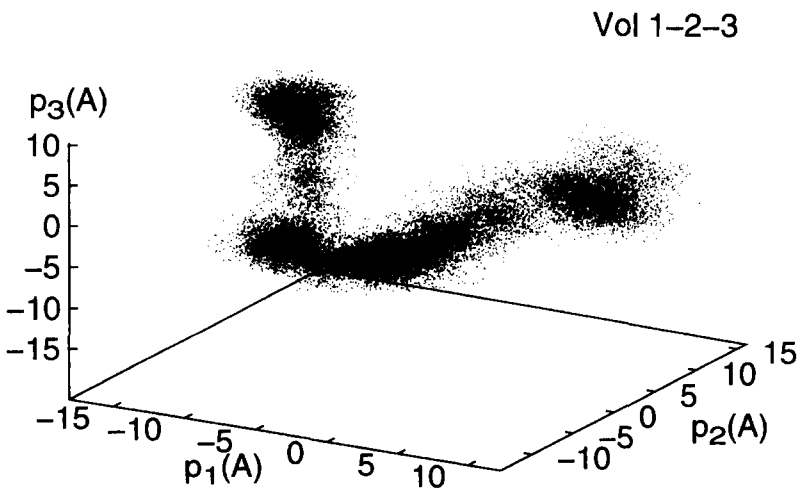

Fig. 5. Projection of the MD trajectory on the three-dimensional subspace spanned by MODC $\boldsymbol{m}_1$, $\boldsymbol{m}_2$ and $\boldsymbol{m}_3$.

3.3. Diffusion in configurational space

With the above formalism we can now describe the MD trajectory of the protein in configurational space as a particle diffusing in real space and getting trapped occasionally in shallow minima for a finite period of time before disentagling itself and moving on. In this section we introduce a model for the dynamics of the protein in which we project this motion only on a subset of the principal MODC that are deemed the relevant coordinates of the system. The protein configuration is considered a stochastic variable, $x(t)$, moving with a random velocity, $\zeta(t)$, in the reduced space of the principal MODC. The MODC are coupled (strongly) to a heat bath consisting of the other (possibly quasi-harmonic) MODC and the solvent. Defining the displacements

$$x(t) = \int_o^t \zeta(t')\,\mathrm{d}t' + x(0). \tag{1}$$

The ensemble average of the square of the mean square displacement (msd) is

$$\langle x^2(t)\rangle = \int_0^t \mathrm{d}t' \int_0^{t'} \mathrm{d}t''\langle\zeta(t')\zeta(t'')\rangle, \tag{2}$$

where we assume no correlations between $x(0)$ and $\zeta(t)$ (i.e., $\int_0^{t'}\langle\zeta(t')x(0)\rangle = 0$). We are looking now for a stationary distribution, namely, for a solution where the velocity–velocity correlation function satisfies

$$\langle\zeta(t')\zeta(t'')\rangle = \langle\zeta(|t'-t''|)\zeta(0)\rangle = \widetilde{C}(|t'-t''|). \tag{3}$$

The time derivative of the msd displacement is

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x^2(t)\rangle = 2\int_0^t \langle\zeta(t')\zeta(0)\rangle\,\mathrm{d}t'. \tag{4}$$

Thus the integrand on the RHS of Eq. (4) is exactly $\widetilde{C}(|t'-t''|)$, the correlation between the velocities of the system at time t and $t+t'$ averaged over a long period of time t. For regular diffusion the RHS of Eq. (4) is a constant which is usually identified as the diffusion coefficient, D. Eq. (4) elucidates the direct relation between the diffusion coefficient and the correlation function, $\widetilde{C}(|t'-t''|)$. When this function decays exponentially or faster for large values of t the integral is finite and the particle (or protein) exhibits a regular diffusive behavior. In contrast, for a system diffusing anomalously $\widetilde{C}(|t'-t''|)$ has a long algebraic tail of the form

$$\widetilde{C}(|t'-t''|) \simeq \kappa/t^{\alpha} \tag{5}$$

(other forms of long tails may also exist but are not addressed here). In such a system the msd is characterized by

$$\langle x^2(t)\rangle \sim t^{2H_D}, \tag{6}$$

where $H_D = 1-\alpha/2$. The exponent H_D is the Hölder exponent, which, in the case of simple Brownian motion, has the value $1/2$. Values of $H_D > 1/2$ ($H_D < 1/2$) correspond to *superdiffusion* (*subdiffusion*). While superdiffusion cannot occur with only partial trapping (which always acts to slow down the tracer particle, not enhance its mobility), occasional very long jumps of the system in the configurational space can give rise to superdiffusion. This sort of behavior belongs to the so-called the 'Lévy flight' class. Inspection of the simulation indicates that indeed such long

jumps can occur, but we have not yet substantiated this result quantitatively and the relevance of the Lévy flight to our system is not clear yet.

In our analysis we define the velocity in the space of the principal MODC as

$$\zeta(i) = \frac{x(i+1) - x(i)}{\Delta t}, \tag{7}$$

where the index i denotes the ith configuration along the trajectory of the system and Δt, the time that it takes for the system to reach $x(i+1)$ from $x(i)$. The (normalized) velocity autocorrelation function is defined via

$$C(t) = \frac{\langle \zeta(t)\zeta(0) \rangle}{\sigma_v^2}, \tag{8}$$

where σ_v^2 is the variance of the velocity distribution. The velocity autocorrelation functions, $C(t)$, along each of the five principal MODC are shown in Fig. 6. Notice that the $C(t)$ approach a small value for times larger than a few (1–2) ps. By determining $\langle |x(t) - x(0)|^2 \rangle$ we can determine the long time behavior of $C(t)$ and $\widetilde{C}(t)$.

Fig. 7 shows log–log plots of the msd, $\langle |x(t) - x(0)|^2 \rangle$, along each of the five principal MODC. From these curves a few features must be pointed out. First, for intermediate times, ranging from 5 to 800 ps, the msd shows a power law with exponent smaller than unity (~ 0.4). Second, the short time ($\sim$ 1–10 ps) behavior is marked by a slower increase than the intermediate range, but the range is too short to determine whether this is a transient, finite size effect or a consistent behavior typical of smaller-scale motions. Third, for all directions, except for the first one, the msd reaches a plateau after a certain time, τ_i, where i stands for the MODC index. The first MODC shows what seems to be a sharper increase, which may indicate a different mechanism for larger configurational

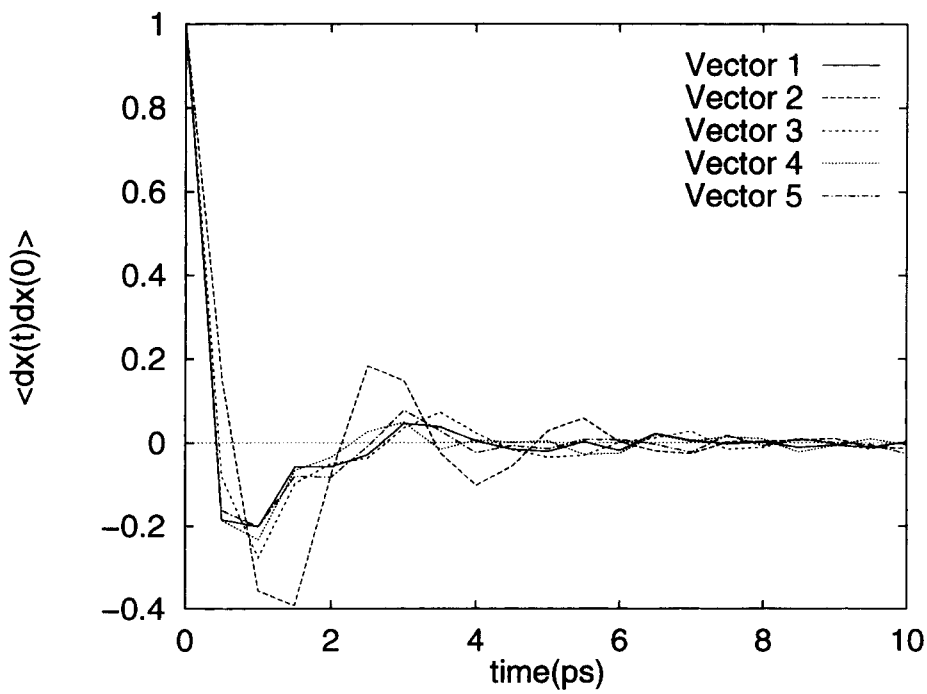

Fig. 6. Velocity autocorrelation function, $C(t)$, along each of the five principal MODC.

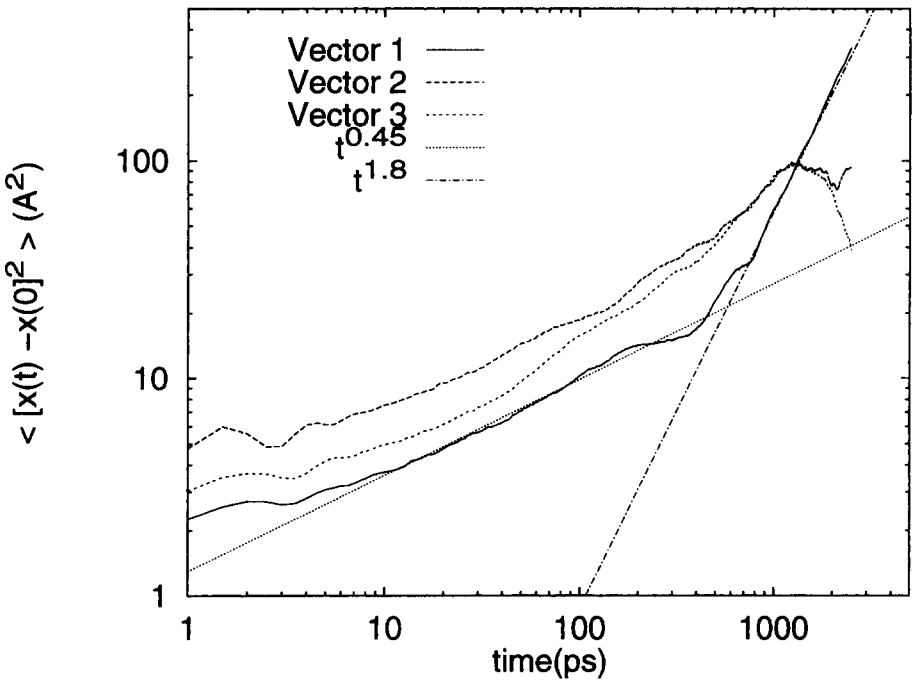

Fig. 7. Mean square displacement along each of the five principal MODC.

transformations. Again, the range is too short to fit this behavior with a good degree of confidence, but a power-law fit would give a value larger than 1.5 for times larger than 2 ns. As a guide to the eye, we have included a line representing a power of 1.8.

The plateau in the msd can be argued to be consistent with finite size cutoffs in self-similar systems. In our application it may reflect the finite length of the system that restricts $\langle |x(t) - x(0)|^2 \rangle$ to be finite. If we consider the projections along each one of the MODC as a one-dimensional random (but not Brownian) walk then our results above indicate that the msd along each of these directions at a given time decreases as one goes from the highest to the lowest eigenvalue. The number of steps in each of these separate one-dimensional systems is the same and hence they can be considered as displaying the same behavior but with different step sizes. This is also consistent with the observation that the crossover to the power law $2H_D = 0.8$ occurs at longer times for higher eigenvalues. The fact that the power law H_D is smaller than $1/2$ implies that the diffusive motion of the system covers less volume in the configurational space than a Brownian motion, indicating a strong suppression of diffusion (subdiffusion). The sharp increase in the first direction indicates a faster spreading along this direction as the protein becomes more rigid and may point to a well-concerted motion. For comparison, if the power at this regime would be $2H_D = 2$, the interpretation would be of an almost ballistic motion ($r \sim t$). On the other hand a systematic increase in the power with time may suggest that we have indeed a combination of partial trapping with a Lévy flight process and an analysis taking the two effects into acount is more appropriate. This will be carried out in a later report.

The msd for the principal MODC (i.e., the one with largest eigenvalue) is unique in not reaching a plateau in our simulation. This suggests that within the 5.1 ns of our simulation the first MODC describes diffusion in a practically *unbounded* region. We expect longer simulation times to exhibit saturation of the msd also along this MODC. The time needed for all directions to reach a plateu can be regarded as the time required for the protein to cover an

appreciable fraction of the configuration space. Similar behavior was observed in a 1.2 ns simulation of crambin in aqueous solution [45].

Using the Fokker–Planck equation, we can define the probability distribution for the stochastic variable $x(t)$ as

$$\frac{\partial P(x,t)}{\partial t} = \left[\int_0^t \widetilde{C}(t')\,\mathrm{d}t'\right] \frac{\partial^2 P(x,t)}{\partial x^2}. \tag{9}$$

Notice that, $\int_0^t C(t')\,\mathrm{d}t' = \mathrm{d}|x(t)-x(0)|^2/\mathrm{d}t \sim t^{2H_D-1}$ for large t. We can use $\lim_{t\to\infty}\langle|x(t)-x(0)|^2\rangle = 2Dt^{2H_D}$ as a scaling law of the msd with time. Defining fractional time $t^* = t^{2H_D}$ we can bypass formalisms involving fractional derivatives and obtain a traditional diffusion equation,

$$\frac{\partial P(x,t^*)}{\partial t^*} = D\frac{\partial^2 P(x,t^*)}{\partial x^2}, \tag{10}$$

whose solution is

$$P(x,t;x',t') = \frac{1}{\sqrt{2\pi Dt^*}}\exp\left[\frac{-(x-x')^2}{4Dt^*}\right] = \frac{1}{\sqrt{2\pi D|t-t'|^{2H_D}}}\exp\left[-\frac{(x-x')^2}{4D|t-t'|^{2H_D}}\right]. \tag{11}$$

The function $P(x,t;x',t')$ describes the conditional probability density of finding the protein in a state $x(t)$ given that it was at $x'(t')$ at a previous time. Note that in fractional time, t^*, the distribution assumes the familiar Gaussian form, while in real time it takes the shape of a stretched exponential, a distribution that is known to go hand in hand with anomalous diffusion, and which has been observed in measurements of protein dynamics [5–8].

4. Conclusions

We have used MD simulation as a convenient method for exploring the configurational dynamics of a small protein in a crystalline environment. Nonlinear motions describing oscillations around multi-centered distributions are responsible for most of the atomic fluctuations sampled by a protein on the timescale of nanoseconds. These atomic fluctuations are not well described by large fluctuations of individual atoms or small groups of atoms, but by concerted motions of many atoms, usually referred to by the term 'collective behavior'. These modes are nonlinear in the sense that they describe stochastic transitions between different basins of attraction. Evidence of these nonlinear modes can be seen in various local structural variables [29] (dihedral angles) and global variables (rms distance between all pairs of configurations and clustering analysis). A method for extracting optimal dynamical coordinates that best describe the protein fluctuations has been presented. A generalization of this method to identify small (1–3) dimensional subspaces of the configurational space has been used to show a description of the protein dynamics within the context of multi-basin dynamics. We have constructed an ultrametric hierarchy that partitions the thermally accessible states into subgroups of states with similar structures, as measured by the rms distance. Using this distance as a measure of the conformational dissimilarities we obtained a set of coordinates (MODC) that best represent the fluctuations of the system. We analyzed the projections of the trajectory along these MODC as a stochastic process and found that the trajectory of the protein in configurational space can be described by an anomalous diffusion process. Assuming partial trapping alone as the cause for the anomalous diffusion, we have constructed a Fokker–Planck equation for the conditional probability density of finding the protein in a state defined by the position $x(t)$ in configurational space given that it was in a state $x'(t')$. We have also discussed a

possible interpretation of the sharp increase of the spreading behavior of the protein along the main axis in terms of a possible Lévy flight picture, where the system's jumps from state to state are more concerted and thus involve reaching further minima in the configurational space at shorter times. A detailed analysis of the occurrence of both partial trapping and a distribution of jumping distances will be reported elsewhere.

Experimental measurements of the rebinding kinetics of CO to myoglobin indeed observed a stretched-exponential time dependence [5–8]. These observations led Frauenfelder et al. [5] to propose the existence of a hierarchy of motions occurring at various timescales, which results from an ensemble of nearly degenerate states separated by a hierarchical distribution of enthalpic energy barriers. This observation is in agreement with our analysis, as manifested in the stretched-exponential form of the probability density $P(x, t; x', t')$ (Eq. (11)). The existence of the hierarchy of substates has already been verified in several studies [1–3]. In analyzing an MD trajectory of crambin in solution [29,44] and in crystals we have seen similar phenomena and have shown that transitions from one basin of attraction to another do not conform to the traditional paradigm of diffusion or Markovian stochastic processes. MD simulations of DNA [50], of a transcription regulation protein [51] (CRP), and of an eleven amino acid [54] (substance P), all show similar behavior and therefore our analysis seems to apply to all these observations, at least qualitatively. The presence of nonlinear excitations in proteins has strong implications on the refinement and interpretation of X-ray crystallographic [55–57] and NMR [58] data.

Acknowledgements

We thank H. Frauenfelder, M. Muthukamar, and Jorge Sobhart for stimulating discussions. We also thank the Working Group on Protein Dynamics at the Center for Nonlinear Studies (CNLS) for their interest in this work, comments and suggestions. This work has been supported by the CNLS at Los Alamos and the US-DOE.

Appendix A

To establish the nature of the conformational space sampled during the MD simulation of a protein we use a set of directions $\boldsymbol{m}^{3N}$ in the $3N$-dimensional conformational space that best (in the least-square sense) describes the structural fluctuations of the molecule under study. The directions $\boldsymbol{m}^{3N}$ are determined by minimizing the mean square distances of the $\{\boldsymbol{r}_i^{3N}\}$ configurations *normal* to $\boldsymbol{m}^{3N}$, such that most of the fluctuations will be along $\boldsymbol{m}^{3N}$. The distance between a point $\boldsymbol{r}_i$, that here represents a biomolecule conformation, and a line with direction $\boldsymbol{m}$, passing through the point $\boldsymbol{y}_0$, is given by

$$d_i^2 = (\boldsymbol{r}_i - \boldsymbol{y}_0)^2 - [(\boldsymbol{r}_i - \boldsymbol{y}_0) \cdot \boldsymbol{m}]^2. \tag{A.1}$$

The average square distance between a set of S points representing all the trajectory points of the biomolecule is then given by

$$d^2 = \frac{1}{S}\sum_{i=1}^{S} d_i^2 = \frac{1}{S}\sum_{i=1}^{S}(\boldsymbol{r}_i - \boldsymbol{y}_0)^2 - [(\boldsymbol{r}_i - \boldsymbol{y}_0) \cdot \boldsymbol{m}]^2. \tag{A.2}$$

The least-square distance is obtained by finding the $6N$ parameters $\boldsymbol{y}_0 = \{y_{0\alpha}\}$, and $\boldsymbol{m} = \{\boldsymbol{m}_\alpha\}$, with $\boldsymbol{m} \cdot \boldsymbol{m} = 1$, that minimize d^2. That is, we have to minimize a functional of the trajectories, $r_i(t)$, and a function of $\boldsymbol{m}$, $\boldsymbol{y}_0$ and λ,

$$f(\boldsymbol{m}, \boldsymbol{y}_0, \lambda) = \frac{1}{S}\sum_{i=1}^{S}\{(\boldsymbol{r}_i - \boldsymbol{y}_0)^2 - [(\boldsymbol{r}_i - \boldsymbol{y}_0)\cdot\boldsymbol{m}]^2\} + \lambda[\boldsymbol{m}\cdot\boldsymbol{m} - 1], \tag{A.3}$$

where λ is a Lagrange multiplier. An extreme value of d^2 is given by a set $z = (m_\alpha, y_{0,\alpha}; \alpha = 1, \ldots, 3N, \lambda)$ that gives $\nabla_z f(z) = 0$. The gradient of $f(\boldsymbol{m}, \boldsymbol{y}_0, \lambda)$ gives:

(i) with respect to $\boldsymbol{y}_0$:

$$\nabla_{\boldsymbol{y}_0} f = \frac{2}{S}\sum_{i=1}^{S}\{-(\boldsymbol{r}_i - \boldsymbol{y}_0) + [(\boldsymbol{r}_i - \boldsymbol{y}_0)\cdot\boldsymbol{m}]\boldsymbol{m}\} = 0 \tag{A.4}$$

that implies $\boldsymbol{y}_0 = \frac{1}{S}\sum_{i=1}^{S}\boldsymbol{r}_i$, i.e., $\boldsymbol{y}_0$ is the average over all configurations;

(ii) with respect to λ: $\nabla_\lambda f = \boldsymbol{m}\cdot\boldsymbol{m} - 1 = 0$, that normalizes the vector $\boldsymbol{m}$;

(iii) with respect to m_α:

$$\nabla_{m_\alpha} f = -\frac{1}{S}\sum_{i=1}^{S}\{(r_i - y_0)_\alpha (r_i - y_0)\cdot\boldsymbol{m}\} + \lambda m_\alpha = 0. \tag{A.5}$$

We can re-write the right-hand side of this equation as

$$\frac{1}{S}\sum_{\beta=1}^{3N}\sum_{i=1}^{S}(r_i - y_0)_\alpha (r_i - y_0)_\beta m_\beta = \lambda m_\alpha. \tag{A.6}$$

Defining

$$\sigma_{\alpha,\beta} = \frac{1}{S}\sum_{i=1}^{S}(r_i - y_0)_\alpha (r_i - y_0)_\beta, \tag{A.7}$$

where $\sigma_{\alpha,\beta}$ is positive semi-definite, we obtain

$$\sigma\cdot\boldsymbol{m} = \lambda\boldsymbol{m}, \tag{A.8}$$

which is an eigenvalue equation for $\sigma\cdot\sigma$ has $3N$ eigenvalues, λ_i, and $3N$ eigenvectors, $\boldsymbol{m}_i$.

To find out the eigenvectors $\boldsymbol{m}_i$ that minimize d^2, we evaluate d^2 for each line defined by the direction $\boldsymbol{m}_i$ and $\boldsymbol{y}_0$, namely,

$$\begin{aligned} d^2(\boldsymbol{m}_k) &= \frac{1}{S}\sum_{i=1}^{S} d_i^2 = \frac{1}{l}\sum_{i=1}^{S}(r_i - y_0)^2 - [(r_i - y_0)\cdot\boldsymbol{m}_k]^2 \\ &= \sum_{\alpha=1}^{3N}\left[\left(\frac{1}{S}\sum_{i=1}^{S}(r_i - y_0)_\alpha^2 - \sum_{\alpha=1,\beta=1}^{3N}(r_i - y_0)_\alpha (r_i - y_0)_\beta \boldsymbol{m}_{k,\alpha}\boldsymbol{m}_{k,\beta}\right)\right] \\ &= Tr(\sigma) - \boldsymbol{m}_k\cdot\sigma\cdot\boldsymbol{m}_k \\ &= Tr(\sigma) - \lambda_k. \end{aligned} \tag{A.9}$$

The eigenvector corresponding to the largest eigenvalue can be regarded as to the direction of the line that passes through the average conformation, $\boldsymbol{y}_0$, which best represents the predominant motions in the protein. The mean square fluctuations are given by $(1/N)Tr(\sigma) = (1/N)\Sigma_i\lambda_i$.

A.1. Projection on higher-dimensional spaces

This formalism can be readily extended to define the best (in the least-square sense) D-dimensional subspace to describe the motions of the protein. We present results for $D = 2$ and 3. This generalization can be carried out by defining the distance of a configurational point, $r_i(t)$, from a D-dimensional subspace as

$$d_i^2 = (r_i - y_0)^2 - \sum_{k=1}^{D} [(r_i - y_0) \cdot \boldsymbol{m}_k]^2, \tag{A.10}$$

where $\boldsymbol{m}_k$ are D vectors spanning the D-dimensional subspace. Then Eq. (A.1) is modified to

$$d^2 = \frac{1}{S} \sum_{i=1}^{S} d_i^2 = \frac{1}{S} \sum_{i=1}^{S} (\boldsymbol{r}_i - \boldsymbol{y}_0)^2 - \sum_{k=1}^{D} [(\boldsymbol{r}_i - \boldsymbol{y}_0) \cdot \boldsymbol{m}_k]^2 \tag{A.11}$$

and Eq. (A.3) is generalized to

$$f(\{\boldsymbol{m}_k\}, \boldsymbol{y}_0, \{\lambda_k\}) = \frac{1}{S} \sum_{i=1}^{S} \left\{ (\boldsymbol{r}_i - \boldsymbol{y}_0)^2 - \sum_{k=1}^{D} [(\boldsymbol{r}_i - \boldsymbol{y}_0) \cdot \boldsymbol{m}_k]^2 \right\} + \lambda_k [\boldsymbol{m}_k \cdot \boldsymbol{m}_k - 1] + \sum_{k,l \neq k}^{D} \lambda_{k,l} (\boldsymbol{m}_k \cdot \boldsymbol{m}_l), \tag{A.12}$$

where λ_k and $\lambda_{k,l}$ are Lagrange multipliers constraining $\boldsymbol{m}_k$ to be orthonormal. Following the procedure leading to Eqs. (A.3)–(A.9), we find that Eq. (A.9) can be generalized to

$$d^2(\{\boldsymbol{m}_k\}) = Tr(\sigma) - \sum_{k=1}^{D} \lambda_k. \tag{A.13}$$

Here $\{\boldsymbol{m}_k\}$ represents any subset of D eigenvectors of σ. This equation shows that the best planes and volumes are spanned by the eigenvectors of σ with the largest two and three eigenvalues, respectively. The fitness of each subspace will depend explicitly on the specific eigenvalues of σ. To use Eqs. (A.9) and (A.13) we need to find the highest eigenvalues and corresponding eigenvectors of σ. Once the eigenvalues and eigenvectors are calculated, the MD trajectory is projected along the eigenvectors, $p_i(t) = r(t) \cdot \boldsymbol{m}_i$, $i = 1, \ldots, D$. Plots of $p_i(t)$ vs. t show the time history (time series) of the trajectory along each direction. Two- and three-dimensional plots of (p_i, p_j) and (p_i, p_j, p_k) show 2D and 3D cuts of the configurational space sampled by the protein. Eigenvectors and eigenvalues are computed from the simulation data by calculating σ in Eq. (A.7).

References

[1] C.W. Rella, A. Kwok, K. Rector, J.R. Hill, H.A. Schewettman, D.D. Dlott and M.D. Fayer, Phys. Rev. Lett. 77 (1996) 1648–1651.
[2] D. Thorn-Leeson and D.A. Wiersma, Nature Struct. Biol. 2 (1995) 848–851.
[3] D. Thorn-Leeson and D.A. Wiersma, Phys. Rev. Lett. 74 (1995) 2138–2141.
[4] S.W. Englander and N.R. Kallenbach, Quarterly Rev. Biophys. 16 (1984) 521.
[5] H.F. Frauenfelder, S.G. Sligar and P.G. Wolynes, Science 254 (1991) 1598–1603.
[6] H. Frauenfelder, P.J. Steinbach and R.D. Young, Chem. Scripta A 29 (1989) 145–150.
[7] A. Ansari, J. Berendzen, S.F. Bowne, H. Frauenfelder, I.E. Iben, T.B. Sauke, E. Shyamsunder and R.D. Young, Proc. Nat. Acad. Sci. USA 82 (1985) 5000–5004.

[8] I.E.T. Iben, D. Braunstein, W. Doster, H. Frauenfelder, M.K. Hong, J.B. Johnson, S. Luck, P. Ormos, A. Schulte, P.J. Steinbach, A.H. Xie and R.D. Young, Phys. Rev. Lett. (USA) 62 (1989) 916–919.
[9] T.R. Sosnick, L. Mayne, R. Hiller and S.W. Englander, Nature Struct. Biol. 1 (1994) 149–156.
[10] T.R. Sosnick, L. Mayne and S.W. Englander, Prot. Struct. Funct. Genet. 24 (1996) 413–426.
[11] G.S. Huang and T.G. Oas, Proc. Nat. Acad. Sci. USA 92 (1995) 6878–6882.
[12] A.R. Fersht, A. Matouschek, M. Bycroft, J.T. Kellis and L. Serrano, Pure Appl. Chem. 63 (1991) 187–194.
[13] C.M. Jones, E.R. Henry, Y. Hu, C.-K. Chan, S.D. Luck, A. Bhuyna, H. Roder, J. Hofrichter and W.A. Eaton, Proc. Nat. Acad. Sci. USA 90 (1993) 11 860–11 864.
[14] T. Pascher, J.P. Chesick, J.R. Winkler and H.B. Gray, Science 271 (1996) 1558–1560.
[15] M.S. Friedrichs, R.A. Goldstein and P.G. Wolynes, J. Mol. Biol. 222 (1991) 1013–1034.
[16] P.E. Leopold, M. Montal and J.N. Onuchic, Proc. Nat. Acad. Sci. USA 89 (1992) 8721–8725.
[17] J.D. Honeycutt and D. Thirumalai, Biopolymers 32 (1992) 695–709.
[18] K.A. Dill, S. Bromberg, K.Z. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas and H.S. Chan, Protein Sci. 4 (1995) 561–602.
[19] A. Šali, E. Shakhnovich and M. Karplus, J. Mol. Biol. 235 (1994) 1614–1636.
[20] N.D. Socci and J.N. Onuchic, J. Chem. Phys. 101 (2) (1994) 1519–1528.
[21] N.D. Socci and J.N. Onuchic, J. Chem. Phys. 103 (11) (1995) 4732–4744.
[22] J.D. Bryngelson, J.N. Onuchic, N.D. Socci and P.G. Wolynes, Prot. Struct. Funct. Genet. 21 (1995) 167–195.
[23] J.N. Onuchic, P.G. Wolynes, Z. Lutheyschulten and N.D. Socci, Proc. Nat. Acad. Sci. USA 92 (1995) 3626–3630.
[24] P.G. Wolynes, J.N. Onuchic and D. Thirumalai, Science 267 (5204) (1995) 1619–1620.
[25] S.E. Radford, C.M. Dobson and P.A. Evans, Nature 358 (1992) 302–307.
[26] S. Khorasanizadeh, I.D. Peters and H. Roder, Nature Struct. Bio. 2 (1995) 193–205.
[27] G.A. Elove, A.K. Bhuyan and H. Roder, Biochemistry 33 (1994) 6925–6935.
[28] J.A. Krumhansl, Proc. Int. Symp. on Computer Analysis for Life Science, Hawashibara Forum 1985, eds. C. Kawabata and A.R. Bishop, Anharmonicity in: Computer Studies of Biopolymers (Ohmsha Ltd., 1986) pp. 78–88.
[29] A. E. García, Phys. Rev. Lett. 68 (1992) 2696.
[30] R.H. Austin, K.W. Beeson, L. Eisenstein, H. Frauenfelder and I.C. Gunsalus. Biochem. 14 (1975) 5355–5373.
[31] I. Schlichting, J. Berendzen, G.N. Phillips and R.M. Sweet, Nature 371 (1994) 808–812.
[32] R. Elber and M. Karplus, Science 235 (1987) 318.
[33] N. Gō and T. Noguti, Chem. Scripta A 29 (1989) 151–164.
[34] R.H. Austin, 1992 Lectures in Complex Systems, eds. I. Nadel and D.L. Stein (1993) pp. 353–400.
[35] P.J. Kraulis, J. Appl. Cryst. 24 (1991) 946–955.
[36] W.A. Hendrickson and M.M. Teeter, Nature 290 (1981) 107.
[37] M.M. Teeter, Proc. Nat. Acad. Sci. USA 81 (1984) 6014–6018.
[38] W.L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. 110 (1988) 1657–1666.
[39] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell and P.A. Kollman, J. Am. Chem. Soc. 117 (1995) 5179–5197.
[40] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey and M.L. Klein, J. Chem. Phys. 79 (1983) 926–935.
[41] T. Darden, D. York and L. Pedersen, J. Chem. Phys. 98 (1993) 10 089–10 092.
[42] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross III, T.E. Cheatam, D.M. Ferguson, U. Chandra Singh, P. Weiner and P.A. Kollman, AMBER, V. 4.1, 1995.
[43] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren and J.R. Kaak, J. Chem. Phys. 81 (1984) 3654.
[44] A.E. García, in: Nonlinear Excitations in Biomolecules Les Editions de Physique, ed. M. Peyrard (Springer, Berlin, 1995) pp. 191–208.
[45] A.E. García, R. Blumenfeld, G. Hummer and J. Sobehart, Proc. 9th Conversation in Biomolecular Stereodynamics, eds. R.H. Sarma and M.H. Sarma (1996) pp. 267–280.
[46] J. McLachlan, J. Mol. Biol. 128 (1979) 49.
[47] M. Levitt, C. Sander and P.S. Stern, J. Mol. Biol 181 (1985) 423.
[48] T. Noguti and N. Gō, Nature 296 (1982) 433.
[49] L. Lebart, A. Morineau and K.M. Warwick, Multivariate Descriptive Statistical Analysis (Wiley, New York, 1984).
[50] A.E. García, D.M. Soumpasis and T.M. Jovin, Biophys. J. 66 (1994) 1742–1755.
[51] A.E. García and J.G. Harman, Protein Sci. 5 (1996) 62–71.
[52] A. Amadei, A.B.M. Linssen and H.C. Berendsen, Proteins Struct. Funct. Genet. 17 (1993) 412–425.
[53] S. Hayward, A. Kitao and N. Gō, Ann. Rev. Phys. Chem. 46 (1995) 223–250.
[54] G. Hummer and A.E. García (1996), submitted.
[55] J.B. Clarage and G.N. Phillips, Acta Cryst. D 50 (1994) 24–36.
[56] J.A. Krumhansl, Proc. in Life Sciences: Protein Structure, Molecular and Electronic Reactivity, ed. R.H. Austin (Springer, New York, NY, 1987).
[57] A.E. García, J.A. Krumhansl and H. Frauenfelder, Proteins Struct. Funct. Genet. 29 (1997) 1–8.
[58] R. Bruschweiler and D.A. Case, Phys. Rev. Lett. 72 (1994) 940–943.