Title: Multi-basin dynamics of a protein in aqueous solution

Author(s): Angel E. Garcia

Submitted to: "Non-linear Excitations in Biomolecules"
Editions de Physique, Michel Peyrard, Editor
Proceedings of the conference held at Centre de Physique des
Houches, France

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Multi-basin dynamics of a protein in aqueous solution

Angel E. García

Theoretical Biology and Biophysics Group, T10, MS K710

Los Alamos National Laboratory,

Los Alamos, New Mexico, 87545, U.S.A.

ABSTRACT

A molecular dynamics simulation of crambin in aqueous solution shows that motions are characteristic of non-linear systems. We describe typical non-linear excitations, such as intermittency, for various representations of the protein dynamics and structure. The protein backbone dihedral angles show fast correlated transitions from one minimum well to another. Each transition is followed by small overdamped oscillations. Equal-time cross correlations of all (ϕ, ψ) angles show that correlations are extended along the backbone chain. An analysis based on a generalized least squares fitting of the protein fluctuations along vectors show that a small set of molecule optimal dynamic coordinates (MODC) describe most of the protein fluctuations. In addition, the MODC describe a trajectory where the protein conformation jumps from one minimum well to another. An extension of the MODC describing 2- and 3- dimensional cuts of the protein configurational space clearly shows a trajectory around multiple basins of attraction.

1. INTRODUCTION

The dynamics of proteins is closely linked to their function. Proteins and nucleic acids exhibit motions with characteristic time-scales ranging from picoseconds (ps) [1, 2] (normal modes of vibrations) to seconds [3] (as studied by hydrogen exchange experiments). Experimental studies on myoglobin suggest the existence of a hierarchy of motions occurring at various time-scales resulting from an ensemble of nearly degenerate states separated by a distribution of enthalpic energy barriers [4-7]. Theoretical evidence (molecular dynamics and Monte Carlo simulations) for the existence of these substates have been reported [8, 9], although the conclusions of these reports are not free of controversy [10]. The dynamics characteristics of such systems has also been described [11].

Theoretical studies of biomolecular dynamics are mostly concentrated on molecular dynamics simulations of sophisticated [12] models where all the biomolecule's atoms and the aqueous solvent are explicitly included in the calculations. This approach has advan-

tages in the sense that many non-linearities in the system are explicitly included in the modeling. A disadvantage of these models is that by following the trajectory of every atom in the system the length of the simulations is limited by the computational resources. Current atomic-level molecular simulations of biomolecules model the dynamics of proteins up to a few nanoseconds [13]. It seems to be a trend in the field that larger systems and longer simulations are attempted. However, the analysis of the extensive wealth of information provided by these simulations is missing.

In this manuscript we will present evidence that shows the presence of multi-basin non-linear motion in proteins in the picoseconds time-scale. A method for extracting vectors that best represent the fluctuations in the systems will be discussed. Recent studies by Flach and Willis [14] have motivated the generalization of this method to find the best planes, volumes, etc., describing the fluctuations of the system. We will show that the molecular dynamics trajectory of the protein is clustered around few local minima (basins of attraction), and that transitions among local minima occur within the 310 ps of the trajectory.

2. DESCRIPTION OF THE SYSTEM

We have studied the dynamics of a small hydrophobic protein, crambin, in aqueous solution, by a molecular dynamics simulation at constant temperature. Crambin is a 46 amino acids amphipatic protein for which high resolution X-ray [15], neutron diffraction [16] and NMR [17-19] data are available. Detailed experimental and theoretical studies of the hydration and dynamics of crambin have been reported in the literature [1, 20-22]. In spite of the detailed experimental and theoretical studies done on crambin, its function is not known.

Crambin is a 46 amino acid protein that contains most structural elements characteristics of larger proteins. Fig.1 shows a ribbon representation of the three dimensional structure of crambin. Starting from the N-terminus and moving along the protein chain we find a β -strand (amino acids 1-4), a loop (amino acids 5-6), a helix (amino acids 7-19), a loop (amino acids 20-22), another helix (amino acids 23-30), another β -strand that makes hydrogen bondings with the first β strand to form a β sheet (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 is not well described by a quasi-one-dimensional chain.

In this simulation study, crambin was contained in a box of dimension $42.11 \times 36.85 \times 29.34$ Å containing 1315 water molecules. The initial conformation of the protein was obtained from the crystallographic coordinates reported by Hendrickson and Teeter [15]. The system contains 4353 atoms; 408 in the protein and 3945 in the solvent. The system was equilibrated during a period of 24 ps. A production period of 310 ps has been obtained. A previous description of the dynamics was reported for the first 216 ps after equilibration [11]. Details about the system and simulation have been described in a previous paper [23]. In the following analyses the dynamics of the protein will be interpreted as that of an open system coupled to an external bath (the solvent) where the protein exchanges energy and momentum with the solvent.

Figure 1: Ribbon representation of the secondary and tertiary structure of crambin. The labels N and C label the amino and carboxy termini, respectively. The labels H1 and H2 show two helical regions.

3. RESULTS AND DISCUSSION

3.1 Localized Non-Linear Motions

Analysis of the distributions of occurrence and the time dependence behavior of the protein backbone dihedral angles [24], (ϕ, ψ) , suggest that the dynamics of crambin is typical of a system with multiple potential energy minima.

Figs. 2a-b show histograms of the occupancy of the backbone ϕ and ψ dihedral angles during the first 216 ps of simulation. The backbone dihedral angles ω remain in a *trans* conformation throughout the simulation. The ϕ and ψ dihedral angles for residues 2-3,33-35 (forming part of β -strands) and 19-21,40-43 (forming parts of *turns*) show bi-modal distributions. The helical regions of the protein, residues 7-18 and 23-30, show sharp unimodal distributions. Time-series of some angles found to have multimodal distributions are shown in Figs. 3 and 4. These time-series are characteristic of systems showing intermittency [25]. That is, there are many fast flips from one conformation to another, with rapid underdamped oscillations between. Figs. 3a-3d show the time dependence of the (ϕ, ψ) angles near the residues Cys(2) and Cys(3) that sample bimodal distributions. The average value for ψ_3 is around 90° , and the standard deviation is around 50° . However, $\psi = 90^\circ$ is hardly ever adopted, and the fluctuations around each minimum are only 15° . Equal times transitions in ψ_3, ϕ_4 , shown in Figs. 3b and 3c, are

Figure 2: Histograms of the fraction of occurrences of the backbone dihedral angles (a) ϕ (on top) and (b) ψ (in degrees) of crambin during 216 ps of simulation. The amino acids sequence is labeled by numbers.

Figure 3: Dihedral angles ψ and ϕ (in degrees) as a function of time (in ps) for the amino acids Cys(4) and Thr(4). Figures are referred as a-e, from top to bottom

anti-correlated. Notice that when ψ_3 changes abruptly from one conformation to another, ϕ_4 also changes abruptly in the opposite sense. Some of these transitions also seem to be coupled to the other two neighboring dihedral angles, ϕ_4 and ψ_3 , shown in Figs. 4a and 4d, but not as strongly as ψ_3 and ϕ_4 . The strongly coupled transition in ψ_3, ϕ_4 will localize the stress induced by one transition to two residues [26]. Similar behavior is exhibited at other backbone positions. For example, Fig.4a-e show the ψ, ϕ angles for residues Pro(19), Gly(20), and Thr(21) that form a *turn*. The same anti-correlations are shown between different angles. However, the transitions are different. Here the coupling between angles is extended to five consecutive ψ and ϕ angles. The ring structure of the proline residue Pro(19) does not allow ϕ_{19} , Fig.4a, to change from g^- to g^+ or t , but two separate values of ϕ_{19} are sampled around g^- . Notice that the pairs of angles (ϕ_{19}, ψ_{19}) , (ψ_{19}, ϕ_{20}) , and (ψ_{20}, ϕ_{21}) are anti-correlated, but (ϕ_{20}, ψ_{20}) are not anti-correlated over the whole trajectory. These motions have the collective effect of changing the *turn* conformation.

The equal-time cross-correlation matrix shows that correlations extend through the chain to non-nearest-neighbor residues, although with diminished cross-correlation values. Fig. 5 shows the equal-time cross-correlation matrix for all (ϕ, ψ) dihedral angles. The matrix elements range from -1.0 (perfect anti-correlation) to 1.0 (perfect correlation). Regions corresponding to the matrix elements are shaded from white (-1.0) to black (+1.0). The cross-correlations matrix elements among nearest-neighbors (elements near the diagonal) are large and negative, showing strong anti-correlation. Large cross-correlation matrix elements among non-nearest-neighbors dihedral angles are obtained for amino acids involved in non-alpha helical regions. Regions of different secondary structure are easily distinguished in the cross-correlation matrix. Helical regions show a strong anti-correlation between neighboring angles, a moderate correlation between next-nearest neighbors, and a weak anti-correlation to amino acids forming a backbone hydrogen bond. These patterns can be easily seen in Fig. 5, near the diagonal, for amino acids 7-15, and

Figure 4: Dihedral angles ψ and ϕ (in degrees) as a function of time (in ps) for the amino acids Pro(19), Gly(20), and Thr(21). Figures are referred as a-e, from top to bottom.

20-28. Non-helical regions show a cross-correlation pattern among non-nearest-neighbor amino-acids that extends over the whole protein. Amino acids that are not directly connected (neither by covalent bonds nor in van der Waals contact) are correlated. These patterns can be seen in Fig. 5, amino acids 16-20 and 31-45.

3.2 Delocalized Non-Linear Motions

The inter-dependence of local structural-variables describing collective, delocalized excitations is not trivial. However, the description of the dynamics of a protein in terms of non-structural variables is desired. To do so we need to find a measure that will represent the fluctuations of the system. We have employed the N -particle root-mean-square (rms) distance [27], $d(t, t')$, between evolving protein configurations. A large rms distance between configurations at short $t - t'$ are indicative of fast configurational changes.

The distance matrix $d(t, t')$ between pairs of conformations at t, t' , sampled every 2 ps, during 310 ps of simulation, is shown in Fig. 6. A darker gray shading implies a larger rms distance between pairs of configurations. A lighter gray shading implies a small rms distance between pairs of configurations. The configurations of the protein during the first 50.0 ps of the trajectory are far away from other configurations in the trajectory. The rms distance smoothly increases from 0 to values near 1.0 Å in a time near 50.0 ps. Oscillations between larger (1.5 - 2.0 Å) and smaller (1.0 Å) rms distances occur also at intervals of about 50 ps. Normal mode analysis of proteins show the lowest frequency modes to have periods in the order of magnitude of a few ps [1, 2]. That is, the motions responsible for these oscillations are not normal modes.

Figure 5: Equal-time cross-correlation matrix among all backbone dihedral angles. Regions representing elements of the matrix are shaded from white (-1.0) to black (+1.0). Contours are drawn to guide the eye toward regions showing correlations (or anti-correlations).

3.2.1 Tree Analysis

The results shown above suggest that the *rms* distance can be used as an effective measure to detect conformational transitions among local minima. The information contained in $d(t, t')$ is enough to build a hierarchical representation of all configurations adopted by the system. 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 [28]:

1. We start with N configurations and a distance matrix, $d(t, t')$, containing the distance among all pairs of configurations. At this stage, each configuration belongs to a separate cluster.
2. We join two distinct ($d \neq 0$) configurations for which $d(t, t')$ is the smallest into one cluster. Now we have $N - 1$ clusters. To build the new distance matrix we take $d^{(N-1) \times (N-1)}(new, k) = \inf[d(i, k), d(j, k)]$. This step is repeated $N - 1$ times, until only

Figure 6: Contour plot of the *root-mean-square* distance between pairs of conformations adopted by the protein every 2 ps along the last 310 ps of the molecular dynamics trajectory. Regions surrounded by the contours are shaded from white ($d \approx 0.50$ Å) to black ($d \geq 2.0$ Å). The largest *rms* distance is 2.45 Å.

one cluster remains.

This hierarchy can be 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. Fig. 7 shows a radial representations of the hierarchy obtained by this algorithm. We have added labels indicating the time (in ps) at which the configuration represented in the tree occurred in the molecular dynamics trajectory. All configurations belong to a cluster with a branch point (labeled O) near the center of the diagram. This point represents the stem of the tree in a hierarchical representation. Each branch emerging from this point represents 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 tree conforms to the ideas presented by H. Frauenfelder [4], where a hierarchy of structures exist and transitions between structures in nearby minima are fast, while transitions to far away states are reached through of multiple jumps to nearby minima. We believe that the tree presented here is just the bottom of this hierarchy; i.e., it goes from structures differing in the position of a few atoms, to structures differing in the relative orientation of helices and turns. The complete hierarchy may extend from folded structures to structures that exhibit completely different folding or unfolded structures. A complete tree may show that the stem represented by O is only one branch of a larger tree. Such a hierarchy cannot be completely sampled by molecular dynamics

Figure 7: Radial tree representation of structures in different clusters. The numbers around the tree show the time (in ps) at which the structure occurred in the trajectory.

simulations of systems that explicitly include all atoms and pretend to follow the history of every molecule in the solvent. The hierarchy presented here can be proven to satisfy ultrametricity. That is, the distance between clusters satisfy $d(i, j) \leq \min[d(i, k), d(j, k)]$.

It is our hypothesis that the oscillations shown in Fig. 6 and the branching of the tree in Fig. 7 represent collective non-linear motions. To show this we will define a set of directions, $\vec{m}$, in the $3N$ dimensional space of the protein, that best represents the mean square fluctuations of the protein atoms around their equilibrium positions. If the oscillations shown in $d(t, t')$ are non-linear excitations, the projections of the protein trajectory along these directions should exhibit a dynamical behavior similar to that shown by the *turn* and β -*strands* dihedral angles. We expect to observe multicentered oscillations, rapid transitions from one center to another, and damped quasi-harmonic oscillations around each center.

3.3 Molecule Optimal Dynamical Coordinates (MODC)

At this point it is desirable to establish the nature of the conformational space sampling (one-basin-of-attraction-quasi-harmonic motion versus multiple-basins-non-linear motions) performed during the molecular dynamics simulation of a protein in solution. The method involves the construction of a set of directions $\vec{m}^{3N}$ in the $3N$ dimensional conformational space that can be most efficiently used to describe the structural fluctuations of the molecule under study. This method has been previously described [11, 33]. A generalization of the method to represent 2-dimensional (planes) and 3-dimensional (volumes) cuts of the (coordinates) configurational space that better represent the dynamics of the system will be presented. These coordinates are specific to the molecule and trajectory

sampled during a molecular dynamics simulation.

The directions $\vec{m}^{3N}$ are determined by minimizing the mean square distances of the $\{\vec{r}_i^{3N}\}$ configurations *normal* to $\vec{m}^{3N}$, such that most of the fluctuations will be along $\vec{m}^{3N}$. The distance between a point $\vec{r}_i$, that here represents a biomolecule conformation, and a line with direction $\vec{m}$, passing through the point $\vec{y}_0$, is given by

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

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

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

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

$$f(\vec{m}, \vec{y}_0, \lambda) = \frac{1}{S} \sum_{i=1}^S \{(\vec{r}_i - \vec{y}_0)^2 - [(\vec{r}_i - \vec{y}_0) \cdot \vec{m}]^2\} + \lambda[\vec{m} \cdot \vec{m} - 1] \quad (2)$$

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

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

$$\nabla_{\vec{y}_0} f = \frac{2}{S} \sum_{i=1}^S \{-(\vec{r}_i - \vec{y}_0) + [(\vec{r}_i - \vec{y}_0) \cdot \vec{m}]\vec{m}\} = 0 \quad (3)$$

that implies

$$\vec{y}_0 = \frac{1}{S} \sum_{i=1}^S \vec{r}_i, \quad (4)$$

That is, $\vec{y}_0$ is the average over all configurations;

ii) with respect to λ :

$$\nabla_\lambda f = \vec{m} \cdot \vec{m} - 1 = 0 \quad (5)$$

that normalizes the vector $\vec{m}$;

iii) with respect to m_α :

$$\nabla_{m_\alpha} f = -\frac{1}{S} \sum_{i=1}^S \{(r_i - y_0)_\alpha (r_i - y_0)_\beta \cdot \vec{m}\} + \lambda m_\alpha = 0. \quad (6)$$

We can re-write the right hand side of Eq. 6 as

$$\frac{1}{S} \sum_{\beta=1}^{3N} \sum_{i=1}^S (r_i - y_0)_\alpha (r_i - y_0)_\beta m_\beta = \lambda m_\alpha. \quad (7)$$

Defining

$$\sigma_{\alpha,\beta} = \frac{1}{S} \sum_{i=1}^S (r_i - y_0)_\alpha (r_i - y_0)_\beta \quad (8)$$

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

$$\sigma \cdot \vec{m} = \lambda \vec{m}, \quad (9)$$

that is the eigenvalue equation for σ . σ has $3N$ eigenvalues, λ_i , and $3N$ eigenvectors, $\vec{m}_i$.

To find out the eigenvectors $\vec{m}_i$ that minimizes d^2 , we evaluate d^2 for each line defined by the direction $\vec{m}_i$ and $\vec{y}_0$. That is,

$$\begin{aligned} d^2(\vec{m}_k) &= \frac{1}{S} \sum_{i=1}^S d_i^2 = \frac{1}{S} \sum_{i=1}^S (r_i - y_0)^2 - [(r_i - y_0) \cdot \vec{m}_k]^2 \\ &= \sum_{\alpha=1}^{3N} \left[\left(\frac{1}{S} \sum_{i=1}^S (r_i - y_0)_\alpha^2 - \sum_{\alpha=1, \beta=1}^{3N} (r_i - y_0)_\alpha (r_i - y_0)_\beta \vec{m}_{k,\alpha} \vec{m}_{k,\beta} \right) \right] \\ &= \text{Tr}(\sigma) - \vec{m}_k \cdot \sigma \cdot \vec{m}_k = \text{Tr}(\sigma) - \lambda_k. \end{aligned} \quad (10)$$

The eigenvector corresponding to the largest eigenvalue, corresponds to the direction of the line passing through the average conformation, $\vec{y}_0$, that best represents the predominant motions in the protein.

Notice that Eqs. 8 and 9 are closely related to the definitions used in the quasi-harmonic approximation [29, 31, 32]. In the quasi-harmonic approximation, the eigenvalue system solved involves the matrix,

$$K_{\alpha,\beta} = kT \sqrt{a_\alpha a_\beta} \sigma_{\alpha,\beta}^{-1}, \quad (11)$$

where σ is defined by Eq. 8, $\sigma_{\alpha,\beta}^{-1}$ refers to an element of the inverse of the matrix σ , and a_α is the mass of atom α . The difference between quasi-harmonic analysis and the analysis presented here is that we do not assume unimodal distributions of the atomic fluctuations (i.e., motions in a single basin of attraction or in other words, around a single minimal energy structure). The quasi-harmonic approximation uses the relation between the mean square displacement and the eigenfrequencies of a harmonic system to identify a set of temperature dependent frequencies. These eigenfrequencies will, under the assumption of harmonicity, determine the thermodynamics of the system in a closed form. The accuracy of the results strongly depend on the assumption of quasi-harmonicity. Any approach that relies on a quadratic form of the Cartesian displacements (i.e., correspondence analysis, [28], quasi-harmonic analysis, etc.) of the molecule will end with the matrix σ , or its inverse. The significance of the eigenvectors and eigenvalues of σ will strongly depend on the model used to interpret them.

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

$$d_i^2 = (r_i - y_0)^2 - \sum_{k=1}^D [(r_i - y_0) \cdot \vec{m}_k]^2, \quad (12)$$

where $\vec{m}_k$ are D vectors spanning the D -dimensional subspace. Then, Eq. 1 can be replaced by

$$d^2 = \frac{1}{S} \sum_{i=1}^S d_i^2 = \frac{1}{S} \sum_{i=1}^S (\vec{r}_i - \vec{y}_0)^2 - \sum_{k=1}^D [(\vec{r}_i - \vec{y}_0) \cdot \vec{m}_k]^2. \quad (13)$$

Figure 8: Projection $p_i(t)$ of the 310 ps molecular dynamics trajectory along the five largest eigenvalues of Eq. 9 are shown on the left-hand-side plots. We refer to the figures as a-e, from top to bottom. The right-hand-side plots show histograms of the frequency of occurrence of all values of $p_i(t)$ for the corresponding vectors. $p_i(t)$ are given in Å, and t in ps. The labels on top of each curve show the eigenvalue ordering (from large to small) and the corresponding eigenvalues, λ (in Å²).

and Eq. 2 is replaced by

$$f(\{\vec{m}_k\}, \vec{y}_0, \{\lambda_k\}) = \frac{1}{S} \sum_{i=1}^S \{(\vec{r}_i - \vec{y}_0)^2 - \sum_{k=1}^D [(\vec{r}_i - \vec{y}_0) \cdot \vec{m}_k]^2\} + \lambda_k [\vec{m}_k \cdot \vec{m}_k - 1] + \sum_{k,l \neq k}^D \lambda_{k,l} (\vec{m}_k \cdot \vec{m}_l), \quad (14)$$

where λ_k and $\lambda_{k,l}$ are Lagrange multipliers constraining $\vec{m}_k$ to be orthonormal. Following the procedure leading to Eqs. 3-10, we find that Eq. 10 can be generalized to

$$d^2(\{\vec{m}_k\}) = Tr(\sigma) - \sum_{k=1}^D \lambda_k. \quad (15)$$

Here $\{\vec{m}_k\}$ represent 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 lowest two and three eigenvalues, respectively. The fitness of each subspace will depend explicitly on the specific eigenvalues of σ . To use Eqs. 10 and 15 we need to find the highest eigenvalues and corresponding eigenvectors of σ . Once the eigenvalues and eigenvectors are calculated, the molecular dynamics trajectory is projected along the eigenvectors,

$$p_i(t) = r(t) \cdot \vec{m}_i. \quad (16)$$

Figure 9: Projection of the 310 ps molecular dynamics trajectory on the plane spanned by directions (a) $\bar{m}_1$ and $\bar{m}_2$, and (b) $\bar{m}_2$ and $\bar{m}_3$.

Plots of $p_i(t)$ vs. t show the 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. 8.

Figs. 8a-e show the projection of the trajectory along the 5 first five directions (left) and the histograms of the occurrence of all values $p_i(t)$ for the same vectors. The histograms of the population distributions can be fitted to multi-centered distributions with three (Figs. 8a-c) clearly distinguishable centers. Each center is indicative of different basins of attraction. A quasi-harmonic approximation is clearly violated for the first three directions. The projection corresponding to the largest eigenvalue, $\lambda = 101.4 \text{ \AA}^2$, shows a trimodal distribution. The distribution centered at $p_1(t) \sim 20 \text{ \AA}$ is only sampled during the first 40 ps of the molecular dynamics trajectory. A transition from that basin to one centered at $p_1(t) \sim -7 \text{ \AA}$ occurs during the next 10-20 ps. Configurations near this basin of attraction are sampled for the next 40 ps. Another fast transition occurs toward the distribution centered at $p_1(t) \sim 3 \text{ \AA}$. Another transition to the distribution with center near $p_1(t) \sim -7 \text{ \AA}$ occur within the next 65 ps. Notice that the transitions from one basin of attraction to another occur fast (less than 5 ps) while the mean time between

Figure 10: Projection $p_i(t)$ of the last 260 ps molecular dynamics trajectory along the five largest eigenvalues of Eq. 9 are shown on the the left-hand-side plots. We refer to the figures as a-e, from top to bottom. The right-hand-side plots show histograms of the frequency of occurrence of all values of $p_i(t)$ for the corresponding vectors. $p_i(t)$ are given in Å, and t in ps. The labels on top of each curve show the eigenvalue ordering (from large to small) and the corresponding eigenvalues, λ (in Å².)

transitions is long (40-70 ps). The nature of the trajectories described here were apparent from the calculation of $d(t, t')$ matrix shown in Fig. 5 and the tree analysis shown in Fig. 6.

Similar patterns of fast inter-basin transitions followed by overdamped oscillations (and possibly transitions to other local minima within each basin of attraction), are observed for the other four directions in Figs. 8b-e. The *rms* fluctuations of the coordinates during the simulation are 0.98 Å, with $tr(\sigma) = 394$ Å². The first five directions describe 60. % of the fluctuations, with the first direction alone describing 25%. Projections of the trajectories on 2-dimensional subspaces of the configurational space will better characterize the nature of the motions described in Fig. 8. Figs. 9a-b show projections of the trajectory on planes spanned by the directions $\vec{m}_1$ and $\vec{m}_2$ (with the largest eigenvalues) and $\vec{m}_2$ and $\vec{m}_3$ (the best planes that exclude the direction $\vec{m}_1$). The initial ($t = 0$ ps) and final ($t = 310$ ps) positions of the trajectory on the planes are labeled in the figures. The distribution of conformations in Fig. 9a show four basins of attraction with centers near $(p_1, p_2) =$ (I) (20,10), (II) (5, -12), (III) (-7,-5) and (IV) (-12,10). These points are chosen to label the four basins and do not carry any other significance. Basin I contains the initial configuration and is well separated from the other three basins. The other three basins are densely sampled during the trajectory. Together with Figs. 8a-b we can

Figure 11: Projection of the last 260 ps molecular dynamics trajectory on the 3-dimensional subspace spanned by directions $\bar{m}_1$, $\bar{m}_2$ and $\bar{m}_3$.

see that the protein samples basin I and undergoes a transition to the other basins. The protein will remain there for the remaining 260 ps of the simulation. Transitions to each of the other basins (II, III, and IV) are observed.

During the last 260 ps of the simulation $p_1(t)$ samples values between -20 and 10 Å, while $p_2(t)$ samples the same range of values as in the whole trajectory. Fig. 9b shows a projection of the trajectory on the plane spanned by the $\bar{m}_2$ and $\bar{m}_3$ directions. The densely sampled distribution of p_2 values is well separated into the third dimension, $\bar{m}_3$. Four densely occupied basins of attraction are observed. A closed pathway interconnecting these basins can also be seen.

The large change in $p_1(t)$ from $p_1 \sim 20$ Å to $p_1 \sim 3$ to -7 Å, and since $p(t)_1 \sim 20$ Å is not sampled again during the simulation lead us to consider this transition as a transient in the dynamics from a metastable initial conformation to a lower-energy conformation. This was first observed in a previous article [11] when a 216 ps trajectory was analyzed. An analysis of a 310 ps trajectory shows the same features. Therefore, we identify this transition as a transient and the first 50 ps of the trajectory are going to be considered as part of the system equilibration. The remaining 260 ps trajectory (after 74 ps equilibration) is analyzed next. Fig. 10 shows the projection of this trajectory along the five largest eigenvalues of Eq. 9. Multicentered distributions of the occurrence of $p_i(t)$ are shown by the first three direction (Figs.10a-c). Patterns of fast inter-basin transitions followed by overdamped oscillations observed in Figs. 8a-e are also present. [It must be clarified that the MODC analysis yields different eigenvalues and eigenvectors when different segments of the trajectory are analyzed.]

The projection of the equilibrated trajectory along the first three directions (i.e., the best 3-dimensional volume) is shown in Fig. 11. Four basins of attraction can be distinguished as they are densely sampled during the trajectory. This volume describes 47 % of

Figure 12: Projection of the last 260 molecular dynamics trajectory on the plane spanned by directions (a) $\vec{m}_1$ and $\vec{m}_2$, and (b) $\vec{m}_1$ and $\vec{m}_3$.

the total fluctuations during the last 260 ps trajectory. Projection of this volumes on the (p_1, p_2) , and (p_1, p_3) planes are shown in Fig. 12. Figs. 12a-b show a clear partitioning of configurational space similar to the one shown in Fig. 9b. The planes (p_1, p_2) and (p_1, p_3) are the two planes that best describe the fluctuations of the system, representing 41% and 37% of the total fluctuations. The projection of the trajectory on the (p_2, p_3) plane (not shown) does not show a clear partitioning of the configurational space. The MODC analysis provides a quantitative tool to dissect 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.

4. CONCLUSIONS

We have used the molecular dynamics technique mainly as a convenient (although not efficient) method of exploring the configurational dynamics of a small protein in solution. From the results presented here we can conclude that non-linear motions describing oscillations around multicentered distributions are responsible for most of the atomic fluc-

tuations sampled by a protein on a 100 ps time-scale. 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. These modes are non-linear in the sense that they describe transitions among different basins of attraction. The mean time between transitions is around 50 ps. The signature of these non-linear modes can be seen in various local structural variables (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 described. A generalization of this method to identify small (1-3) dimensional subspaces of the configurational space has been used to show, for the first time, a conclusive description of the protein dynamics within the context of multi-basin dynamics.

Acknowledgements

I wish to offer special thanks to Prof. J. A. Krumhansl for his interest on this work, comments and suggestions. I also want to thank Drs. G. Hummer, Javier Sobehart, and Raphael Blumfield, members of the Los Alamos National Laboratory (LANL) Center for Non-Linear Studies (CNLS) Protein Dynamics Working Group (PDWG) for comments and suggestions. This work has been supported by the US-DOE under LANL LDRD-PD research funds.

References

- [1] Levitt, M., C. Sander, and P.S. Stern, *J. Mol. Biol.*, **181**, 423 (1985).
- [2] Noguti, T. and Go, N., *Nature*, **296**, 433 (1982);
- [3] Englander, S.W., and Kallenbach, N.R., *Quarterly Rev. Biophys.*, **16**, 521 (1984).
- [4] Frauenfelder, H., Siglar, H.A., and Wolynes, P., *Science*, **254**, 1598 (1991)
- [5] Frauenfelder, H., Steinbach, P.J., and Young, R.D., *Chemica Scripta*, **29A**, 145 (1989).
- [6] Ansari, A. et al., *Proc. Natl. Acad. Sci. (USA)*, **82**, 5000 (1985)
- [7] Iben, I.E.T. et al., *Phys. Rev. Lett*, **62**, 1916 (1989).
- [8] Elber, R. and Karplus, M., *Science*, **235**, 318 (1987)
- [9] Go, N. and Noguti, T., *Chemica Scripta*, **29A**, 151 (1989).
- [10] Austin, R.H., in *1992 Lectures in Complex Systems*, L. Nadel and D.L. Stein, Editors. (Addison-Wesley, New York. 1993.) pp. 353-400.
- [11] García, A.E. *Phys. Rev. Lett.* **68**, 2696 (1992).
- [12] Webster Dictionary defines *sophisticatus* : aj [ML *sophisticatus*] 1: not in a natural, pure, or original state : Adulterated {a oil} 2: deprived of native or original simplicity : as 2a: highly complicated : Complex { instruments} 2b: Worldly-wise,

Knowing {a adolescent} 3: devoid of grossness : Subtle : as 3a: finely experienced and aware {a columnist} 3b: intellectually appealing {a novel} - so.phis.ti.cat.ed.ly av.

- [13] Steinbach, P.J. and B.A. *Proc. Natl. Acad. Sci. (USA)*, **90**, 9135 (1994).
- [14] Flach, S. and Willis, C.R., *Phys. Rev. Lett.*, **72**, 1777 (1994).
- [15] Hendrickson, W.A. and M.M. Teeter, *Nature*, **290**, 107 (1981).
- [16] Teeter, M.M., *Proc. Natl. Acad. Sci. (USA)*, **81**, 6014 (1984).
- [17] Llinás, M., A. de Marco, and J.T.J. Lecomte, *Biochem.*, **19**, 1140 (1980).
- [18] Usha, M.G. and R.J. Wittebort, *J. Mol. Biol.*, **208**, 669 (1989).
- [19] Vermeulen, J.A.W.H., R.M.J.N. Lamericks, L.J. Berliner, A. de Marco, M. Llinás, R. Boelens, J. Alleman, and R. Kaptein, *Febs*, **219**, 426 (1987). R.M.J.N. Lamericks, L.J. Berliner, R. Boelens, A. de Marco, M. Llinás, and R. Kaptein, *Eur. J. Biochem.*, **171**, 307 (1988).
- [20] Whitlow, M. and M.M. Teeter, *J. Amer. Chem. Soc.*, **108**, 7163 (1986).
- [21] Teeter, M.M. *Ann. Rev. Biophys. Biophys. Chem.*, **20**, 577 (1991).
- [22] Teeter, M.M. and D.A. Case, *J. Phys. Chem.*, **94**, 8091 (1990).
- [23] García, A.E. & Stiller, L. *J. Comp. Chem.*, **12**, 1 (1993).
- [24] Schulz, G.E. and R.H. Schirmer, *Principles of Protein Structure*, (Springer-Verlag, NY, 1978.)
- [25] Krumhansl, J.A., in *Computer Analysis for Life Science*, C. Kawabata and A.R. Bishop, Eds., (Ohmsha, LTD, Tokyo, Japan, 1985.) pp. 78-88.
- [26] Krumhansl, J.A., in *Proceedings in Life Sciences: Protein Structure, Molecular and Electronic Reactivity*, R.H. Austin, et al., Eds., (Springer Verlag, New York, 1987.)
- [27] McLachalan, J., *J. Mol. Biol.*, **128**, 49 (1979).
- [28] Lebart, L., Morineau, A., and Warwick, K. M., *Multivariate Descriptive Statistical Analysis*, (John Wiley & Sons, New York, 1984).
- [29] Karplus, M. and Kushick, J.N., *Macromolecules*, **14**, 325 (1981);
- [30] Ichiye, T. and Karplus, M., *Biochemistry*, **27**, 3487 (1988).
- [31] *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*. Brooks, C.L., Karplus, M. & Montgomery-Pettitt, B., *Advances in Chemical Physics*, **LXXI**, (John Wiley & Sons, New York, 1988.)
- [32] Levy, R.M., Karplus, M., Kushick, J.N. and Perahia, D., *Macromolecules*, **17**, 1370 (1984).
- [33] García, A.E., Soumpasis, D.M. and Jovin, T.M., *Biophys. J.*, **66**, 1742 (1994).