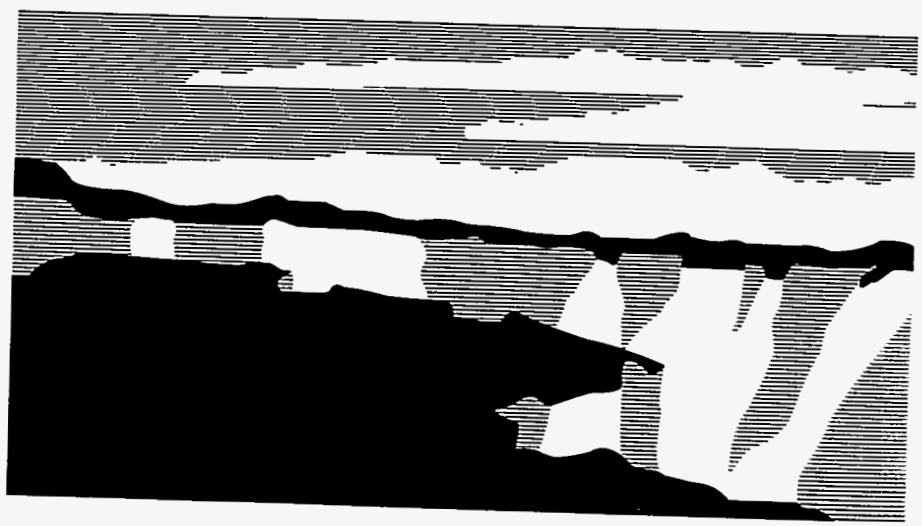
CONF-9506267--1

LA-UR- 95- 2962

Title: Diffusion of a Protein in Configuration Space

Author(s): Angel E. Garcia, Raphael Blumenfeld, Gerhard Hummer, and Jorge Sobehart

Submitted to: Proceeding of the 9th Conversation in Biomolecular Stereodynamics, June 20-24, 1995, Albany, NY



Los Alamos
NATIONAL LABORATORY

Form No. 836 R5
ST 2629 10/91

Diffusion of a protein in configuration space

Angel E. García[1],Raphael Blumenfeld[2],

Gerhard Hummer[1,2], and Jorge Sobehart[2,3]

[1]Theoretical Biology and Biophysics Group, T10, MS K710

[2]Center for Nonlinear Studies

and [3]CIC-3

Los Alamos National Laboratory,

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

in: *Proceedings of the 9th Conversation in Biomolecular Stereodynamics*
edited by R.H. Sarma, (Adenine Press, Schenectady, NY, 1996)

Abstract

Simulations of biomolecular dynamics are commonly interpreted in terms of harmonic or quasi-harmonic models for the dynamics of the system. These models assume that biomolecules exhibit oscillations around a single energy minimum. However, spectroscopic data on myoglobin suggest that proteins sample multiple minima. Transitions between minima reveal a broad distribution of energy barriers. This behavior has been observed in other biomolecular systems.

To elucidate the nature of protein dynamics we have studied a 1.2ns molecular dynamics trajectory of crambin in aqueous solution. This trajectory samples multiple local energy minima. Transitions between minima involve collective motions of amino acids over long distances. We show that nonlinear motions are responsible for most of the atomic fluctuations of the protein. These atomic fluctuations are not well described by large motions of individual atoms or a small group of atoms, but rather by concerted motions of many atoms. These nonlinear motions describe transitions between different basins of attraction. The signature of these motions manifests in local and global structural variables.

A method for extracting Molecule Optimal Dynamic Coordinates (MODC) is presented. A generalization of this method is used to identify small (1-3) dimensional subspaces of the configuration space

describe the dynamics of the protein within the context of nonlinear, multi-basin system.

We present a model for describing the dynamics of biomolecules in terms of an open Newtonian system (protein) coupled to a stochastic system (solvent). Autocorrelation functions of the displacements along relevant MODC show that the protein loses *memory* of its configuration within a few picoseconds. The diffusion of the protein in configuration space is anomalous, namely, the time dependence of the mean square displacement is not proportional to time, but to t^{2H_D} where $2H_D$ is a nontrivial fractional exponent. Therefore, transitions among energy minima far apart in configuration space exhibit a stretched-exponential time dependence, scaling as $t^{-2H_D}\exp(-t^{-2H_D})$, with $H_D < 0.5$. This picture is consistent with a model suggested by Frauenfelder and collaborators to explain multiple timescale relaxation processes observed in myoglobin.

1 INTRODUCTION

Experimental studies of the low temperature ($T < 180$K) rebinding kinetics of CO and O_2 to myoglobin exhibit a stretched-exponential time dependence (2). This suggests 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 (1, 7, 8, 15). This behavior is not unique to myoglobin and it seems to be a characteristic of many biomolecular systems (23). However, this does not imply that

this hierarchy of substates is always directly involved in protein function. Numerical evidence (molecular dynamics and Monte Carlo simulations) for the existence of these substates have been reported (6, 13), although the conclusions of these reports are not free of controversy (3). The dynamic characteristics of such systems has also been described (9, 12).

Here we demonstrate the existence of multi-basin nonlinear motions in proteins in the picosecond time-scale. We will describe a method for extracting coordinates that best represent the fluctuations in the system. We show that the molecular dynamics trajectory of the protein is clustered around few local minima (basins of attraction), and that many transitions among local minima occur within the 1.2 ns trajectory.

We put forward an alternative model for describing the dynamics of biomolecules in which the trajectory is analyzed in terms of an open system (describing relevant dynamical variables of the protein) coupled to a heat bath (i.e., solvent and other, less relevant, protein degrees of freedom). This analysis shows that the protein loses *memory* of its configuration within a few picoseconds. We find that the diffusion of the protein state in configuration space is anomalous. That is, the time dependence of the mean square displacement (*msd*) is not proportional to time, but exhibits a power law behavior, t^{2H_D}. This implies that transitions among energy minima far apart in configuration space will exhibit a stretched-exponential time dependence, $\sim t^{-2H_D} \exp[t^{-2H_d}]$.

The analysis of an MD trajectory presented here is consistent with a model suggested by Frauenfelder and collaborators (1, 2, 7, 8, 15) to explain multiple timescale relaxation times observed in myoglobin. We show here that the protein samples multiple local energy minima that can be classified into an ultrametric hierarchy. Thus we calculate that the diffusion of the protein in configuration space can be characterized by a stretched-exponential in time.

2 DESCRIPTION OF THE SYSTEM

We have studied the dynamics of a small hydrophobic protein, crambin, in aqueous solution, by a molecular dynamics (MD) simulation at constant temperature. Crambin is a 46 amino acids amphipathic protein for which high resolution X-ray (14), neutron diffraction (24) and NMR (20, 27, 28) data are available. Detailed experimental and theoretical studies of the hydration and dynamics of crambin have been reported in the literature (19, 25, 26, 29). Crambin is a small that contains most structural elements characteristic of larger proteins. 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), another loop (amino acids 20-22), another helix (amino acids 23-30), another β-strand that makes hydrogen bond 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 $42.11 \times 36.85 \times 29.34$ $\AA^3$ containing 1315 water molecules. The initial conformation of the protein was obtained from the crystallographic coordinates reported by Hendrickson and Teeter (14). The system contains 4353 atoms; 408 in the protein and 3945 in the solvent. The system was equilibrated during a period of 24 ps. The production extended over 1200 ps. A previous description of the dynamics was reported for the first 216 ps after equilibration (9). Details about the system and simulation have been described in a previous paper (10).

3 *RESULTS AND DISCUSSION*

3.1 Distance Matrix

In a previous paper we have shown that the distributions and time dependence of the protein backbone dihedral angles (22), $(\phi,\ \psi)$, are typical of a system with multiple potential energy minima (9). The ϕ and ψ dihedral angles for residues forming part of β-strands and *turns* show bi-modal distributions while the helical regions of the protein show sharp unimodal distributions. Time-series of some angles are found to be characteristic to systems showing intermittency (17). That is, there occur many fast flips from one conformation to another, following rapid underdamped oscillations.

The inter-dependence of local structural-variables describing collective, delocalized excitations is not trivial. A the description of the dynamics of a protein in terms of non-structural variables is desired. To obtain such a description 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 (21), $d(t,t')$, between evolving protein configurations. A large *rms* distance between configurations at short $t-t'$ are indicative of fast configuration changes.

The distance matrix $d(t,t')$ between pairs of conformations at t,t', sampled every 6 ps, during the simulation, is shown in Fig. 1. A darker gray shading implies a large *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 ps of the trajectory are far away from other configurations in the trajectory. The *rms* distance smoothly increases from zero to about one Å in a time near 50 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 of a few ps (19), indicating that these oscillations are *not* normal modes.

3.1.1 Tree–Analysis. Classification of Sampled Conformations

The results shown above suggest that the *rms* distance can be used to detect conformational transitions among local minima. The information contained in $d(t,t')$ is sufficient 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 (18): *First*, start with N 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. *Second*, join two distinct ($d \neq 0$) configurations, i and j, 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 $d^{(N-1)\times(N-1)}(new,k) = \min[d(i,k),d(j,k)]$, where k runs over the structures in all remaining clusters. This step is repeated $N-1$ 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. This hierarchy can be indexed by the distance between clusters. Fig. 2 shows a radial representations of the hierarchy obtained by clustering configurations sampled at constant time intervals (3 ps) along the trajectory. 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 (7), where a hierarchy of structures exist and transitions between structures in nearby minima are fast, while transitions to far away states are reached through multiple jumps. 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 struc-

tures that exhibit completely different folding or unfolded structures. It is quite plausible that a complete tree may show that the stem represented by O is only one branch of a larger tree. The hierarchy presented here have been constructed to satisfy *ultrametricity* (i.e., the distance between clusters satisfy $d(i,j) \leq \min[d(i,k), d(j,k)]$, for all k). However, ultrametricity is a consequence of the choice made for the distance between clusters in the second step of the clustering algorithm and do not fully reflect properties of the biomolecular system.

3.2 Molecule Optimal Dynamical Coordinates (MODC)

3.2.1 Method

The oscillations shown in Fig. 1 and the branching of the tree in Fig. 2 are the signature of collective nonlinear motions (9, 12). It is important to establish the nature of the conformational space sampling (single-basin, i.e.; quasi-harmonic motion versus multi-basin nonlinear motions) performed during the molecular dynamics simulation of a protein in solution. To do this we use a method that involves the construction of a set of directions $\vec{m}$ in the 3N dimensional conformation space that systematically describes the structural fluctuations of the molecule under study. This method has been described previously (9, 11, 12). A generalization of the method to represent two-dimensional (plane) and three-dimensional (volume) cuts of the configuration space that best represent the dynamics of the system has also been published (12). These coordinates are specific to the molecule and trajectory sampled during a molecular dynamics simulation.

The directions $\vec{m}$ are determined by minimizing the mean square distances of the $\{\vec{r}_i\}$ configurations *normal* to $\vec{m}$, such that most of the fluctuations will be along $\vec{m}$. The distance between a point $\vec{r}_i$, that here represents a given biomolecule conformation, and a line parallel to $\vec{m}$, passing through a 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 points along the trajectory of the biomolecule is then given by:

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

The least square distance is obtained by finding the parameters $\vec{y}_0 = \{y_{0\alpha}\}$, and $\vec{m} = \{\vec{m}_\alpha\}$, 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,

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

where λ is a Lagrange multiplier. Mimization of Eq.(2) gives:

$$\vec{y}_0 = \frac{1}{S}\sum_{i=1}^{S}\vec{r}_i \quad , \qquad [3]$$

indicating that $\vec{y}_0$ is the average over all configurations, and

$$\sigma\cdot\vec{m} = \lambda\vec{m}, \qquad [4]$$

where

$$\sigma_{\alpha\beta} = \frac{1}{S}\sum_{i=1}^{S}(r_i - y_0)_\alpha(r_i - y_0)_\beta. \qquad [5]$$

Here σ has 3N eigenvalues, λ_i, and 3N eigenvectors, $\vec{m}_i$.

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

$$d^2(\vec{m}_i) = Tr(\sigma) - \lambda_i \quad . \qquad [6]$$

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

Eqs. (4) and (5) are closely related to the definitions used in the quasi-harmonic approximation (4, 16), in which the eigenvalue system solved involves the matrix

$$K_{\alpha\beta} = kT\sqrt{a_\alpha a_\beta}\,\sigma^{-1}_{\alpha\beta} \; . \qquad [7]$$

Here σ is defined by Eq. (5) , $\sigma^{-1}_{\alpha\beta}$ 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 ore in other words, around a single minimum energy structure). The quasi-harmonic approximation assumes 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 resulting thermodynamics strongly depends on the assumption of quasi-harmonicity, which is incorrect. Clarage et al. (5) have incorrectly cited our work (9) as to imply just the opposite. Any approach that relies on a quadratic form of the Cartesian displacements (i.e., correspondence analysis or principal component analysis (18), quasi-harmonic analysis, etc.) of the molecule will end with either the matrix σ or its inverse. The similarity in the mathematics is not a reflection of the diagonally opposite positions adopted in interpreting the results. The significance of the eigenvectors and eigenvalues of σ will strongly depend on the model used to interpret them. In any case, the time series and distribution of the projection along the MODC are clearly indicative of nonlinear dynamics.

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 (12). This generalization gives

$$d^2(\{\vec{m}_k\}) = Tr(\sigma) - \sum_{k=1}^{D} \lambda_k \ . \qquad [8]$$

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 largest two and three eigenvalues, respectively. The fitness of each subspace will depend explicitly on the specific eigenvalues of σ. Depending on the number of dominant largest eigenvalues in this Eq. (8), we can now choose a small number of representative coordinates (typically two or three) to describe the dynamics of the system.

Once the eigenvalues and eigenvectors are calculated, the molecular dy-

namics trajectory is projected along these principal coordinates

$$p_i(t) = r(t) \cdot \vec{m}_i. \qquad [9]$$

Plots of $p_i(t)$ *versus* t in Fig. 3 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) (Figs. 4 and 5, respectively) show 2D and 3D cuts of the configuration space sampled by the protein. Eigenvectors and eigenvalues are computed from the simulation data by calculating σ in Eq. (5).

3.2.2 Numerical Results

Figs. 3a-e show the projection of the trajectory along the five principal MODC (left) and the histograms of the occurrence of all values $p_i(t)$ for the same coordinates. The histograms of the population distributions can be fitted to multi-centered distributions. Each center is indicative of different basins of attraction. The time series resulting from the trajectory projections along the MODC are also characteristic of nonlinear systems. Patterns of fast inter-basin transitions followed by overdamped oscillations (and possibly transitions to other local minima within each basin of attraction) are observed. The *rms* fluctuations of the coordinates during the simulation are 1.38 Å, with $Tr(\sigma) = 779.3$ Å^2. The first five directions account for 73% of the fluctuations (as measured by the *msd*), with the first direction alone accounting for 43%.

Projections of the trajectories on 2-dimensional subspaces of the configuration space better characterize the nature of the motions described in Fig. 3. Figs. 4a-b show projections of the first 310 ps 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 this trajectory on the planes are labeled in the figures. The distribution of conformations in Fig. 4a 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 identify 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.

Figs. 5a-b show the projections of the trajectories on two-dimensional subspaces spanned by directions $\vec{m}_1$ and $\vec{m}_2$ obtained after analyzing 0.75 ns and a 1.2 ns trajectories. The initial and final points (in time) in the trajectories are labeled by I and F, respectively. These two-dimensional projections of the trajectories are *closed*, namely, basins of attraction are revisited. The sampled configuration space describes a torus. The basins sampled during the 310 ps trajectory (shown in Fig. 4b) are contained within the lower left quadrant of Fig. 5a. However, at the larger time and length scales shown in Fig. 5a we can only distinguish two basins of attraction. The four basins shown in 4b are within one basin in Fig. 5a. This illustrates the self-similarity of the trajectory. When comparing Figs. 5a and 5b notice that the directions $\vec{m}_1$ and $\vec{m}_2$ are exchanged (i.e., the figure is rotated) and that a new region of configuration space along $\vec{m}_1$ was sampled during the last 0.4 ns of the 1.2 ns trajectory.

The trajectory projected on a three-dimensional subspace spanned by the first three MODC is not closed. That is, basins that appear to be resampled in a two dimensional projection are not resampled when viewed in a 3-dimensional projection. Therefore, conclusions regarding the equilibration of the system must be judged depending on the criteria used for defining *equilibration.* In the field of biomolecular dynamics it is customary to interpret a plateau in the *rms* distance from the initial structure, $d(t, t' = 0)$, as a signal of equilibration. Fig. 1 clearly shows that this is not the case. The *rms* distances among structures separated in time reaches a plateau, but this only implies that they are sampling different conformations. We have used the projections along MODC to look for resampling of configurations along the trajectory. One-dimensional projections show resampling of basins within short (200 ps) timescales (9). Two-dimensional projections show resampling of basins within longer (750 ps) timescale, and three-dimensional projections do not show resampling of basins within the 1.2 ns trajectory.

3.3 Diffusion in Configuration Space

In this section we introduce a model for the dynamics of the protein in which a subset of the principal MODC is considered as the relevant coordinates of the system. The protein configuration is considered a stochastic variable, $\zeta(t)$, moving in the reduced space of the principal MODC. It is coupled (strongly) to a heat bath consisting of the other MODC and the solvent. Defining

$$x(t) = \int_o^t \zeta(t')dt' + x(0) \qquad [10]$$

The ensemble average of the square of the mean square displacement is

$$\langle x^2(t)\rangle = \int_0^t dt' \int_0^{t'} dt'' \langle x(t')x(t'')\rangle \qquad [11]$$

Assuming no correlations between $x(0)$ and $\zeta(t)$ we have $\int_0^{t'}\langle\zeta(t')x(0)\rangle = 0$. We are looking for a stationary distribution, namely, for a solution where

$$\langle\zeta(t')\zeta(t'')\rangle = \langle\zeta(|t'-t''|)\zeta(0)\rangle = \tilde{C}(|t'-t''|) \qquad [12]$$

Thus the time derivative of the displacement is

$$\frac{d}{dt}\langle x^2(t)\rangle = 2\int_0^t \langle\zeta(t')\zeta(0)\rangle dt' \qquad [13]$$

The integrand on the r.h.s. of Eq.(13) is $\tilde{C}(|t'-t''|)$, the correlation between the position of the system at time t and the position at time $t+t'$. For regular diffusion the r.h.s. of Eq. (13) is constant which is usually identified as the diffusion constant, D. Eq. (13) elucidates the direct relation between the diffusion coefficient and the correlation function, $\tilde{C}(|t'-t''|)$. When this function decays exponentially for large values of t the integral is finite (regular diffusion). In contrast, for a system diffusing anomalously $\tilde{C}(|t'-t''|)$ has long algebraic tails

$$\tilde{C}(|t'-t''|) \simeq \kappa/t^{\alpha}. \qquad [14]$$

In such a system the mean square of the displacement is characterized by

$$\langle x^2(t)\rangle \sim t^{2H_D}, \qquad [15]$$

which identifies H_D as

$$H_D = 1 - \alpha/2. \qquad [16]$$

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).

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

$$v(t) = \frac{x(i+1) - x(i)}{\Delta t} = \frac{\zeta(i)}{\Delta t} \qquad [17]$$

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

$$C(t) = \frac{\langle v(t)v(0)\rangle}{\sigma_v^2}, \qquad [18]$$

where σ_v^2 is the variance of the velocity distribution. The velocities autocorrelation function, $C(t)$, along each of the five principal MODC are shown in Fig. 6. Notice that C(t) approaches a small value for times larger than a few (2–3) 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, the short time behavior exhibits a power law in time with an exponent larger than one. Second, for intermediate times, ranging from 1 -100 ps, the *msd* shows a power law with exponent smaller than one (~ 0.8). The later value is indicative of anomalous subdiffusion. Third, for all directions, except for the first one, the *msd* reached a plateau after a certain time, τ_i, where i stands for the MODC index.

This behavior is consistent with finite size cutoffs in self similar systems. In our application it reflects the finite length of the system that forces $\langle|x(t) - x(0)|^2\rangle$ to be finite. If we consider the projections along each one of the MODC as a one-dimensional random (but no Brownian) walk then our results above indicate that the square displacement along each of these directions also decreases 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 $H_D = 0.4$ is smaller than 1/2 indicates that the diffusive motion of the system covers less volume in the configurations space than a Brownian motion, namely, subdiffusion.

The *msd* for the principal (i.e., the one with largest eigenvalue) MODC does not reach a plateau in our simulation. This implies that within the 1.2 ns of our simulation the first MODC describes diffusion in an unbounded region. Longer simulation times will allow the *msd* along this MODC to also reach a plateau. The time needed for all directions to reach a plateu in the *msd* displacement can be taken as the time required for the system to appropriately cover configuration space.

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

$$\frac{\partial P(x,t)}{\partial t} = [\int_0^t \tilde{C}(t')dt'] \frac{\partial^2 P(x,t)}{\partial x^2}. \qquad [19]$$

Notice that, $\int_0^t C(t')dt' = d|x(t) - x(0)|^2/dt \sim t^{2H_D-1}$, 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 get the diffusion equation

$$\frac{\partial P(x,t^*)}{\partial t^*} = D\frac{\partial^2 P(x,t^*)}{\partial x^2}, \qquad [20]$$

with solution

$$\begin{aligned} P(x,t;x',t') &= \frac{1}{\sqrt{2\pi D t^*}}\exp[\frac{-(x-x')^2}{4Dt^*}] \\ &= \frac{1}{\sqrt{2\pi D|t-t'|^{2H_D}}}\exp[-\frac{(x-x')^2}{4D|t-t'|^{2H_D}}] \end{aligned} \qquad [21]$$

For long times, the time dependence of the transition of the protein from one state characterized by the variable x' to another state characterized by x, obeys a stretched-exponential. In the experiments on myoglobin by Frauenfelder et al., (1, 2, 7, 8, 15) the state x' is the state immediately after photolysis and the trajectory $x(t)$ describes the rebinding of CO. In crambin, the MODC represent events that influence the Cartesian-coordinate fluctuations and not a reaction coordinate. However, by identifying a signal that characterizes two states along the MODC we expect to observe a similar stretched-exponential time dependence.

4 CONCLUSIONS

We have used molecular dynamics as a convenient method of exploring the configuration dynamics of a small protein in solution at room temperature. From the results presented here we can conclude that nonlinear motions describing oscillations around multicentered distributions are responsible for most of the atomic fluctuations 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 nonlinear in the sense that they describe transitions between different basins of attraction. Evidence of these nonlinear 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 better describe the protein fluctuations has been presented. A generalization of this method to identify small (1-3) dimensional subspaces of the configuration space has been used to show a description of the protein dynamics within the context of multi-basin dynamics.

Experimental measurements of the rebinding kinetics of CO to myoglobin observed a stretched-exponential time dependence (1, 7, 8, 15). These observations led Frauenfelder et al. (7) 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. In analyzing an MD trajectory of crambin in solution we have seen essentially the same phenomena and have shown that nonlinear motions are responsible for transitions from one basin of attraction to another. 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 the *rms* distance as a measure of the conformational dissimilarities we obtained a set of coordinates (MODC) that best represent the fluctuations of the system. The treatment of the projections of the trajectory along these MODC as stochastic variables leads to a description of the trajectory of the protein in configuration space as an anomalous diffusion process. 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 $x'(t')$. The solution to this equation exhibits a stretched-exponential in time, consistent with experimental observations in myoglobin.

Acknowledgements

We wish to offer special thanks to James A. Krumhansl, Hans Frauenfelder and the working group on protein dynamics at the Center for Nonlinear Studies for their interest in this work, comments and suggestions. This work has been supported by the US-DOE under LANL LDRD-PD research funds.

References and Footnotes

1. Ansari, A. et al., *Proc. Natl. Acad. Sci. (USA),* **82**, 5000 (1985)
2. Austin, R.H., Beeson, K.W., Eisenstein, L., Frauenfelder, H., and Gunsalus, I.C., *Biochem.* **14**, 5355-5373 (1975).
3. Austin, R.H., in *1992 Lectures in Complex Systems*, L. Nadel and D.L. Stein, Editors. (Addison-Wesley, New York. 1993.) pp. 353-400.
4. *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics.* Brooks, C.L., Karplus, M. & Montgomery-Pettitt, B., *Advances in Chemical Physics,* **Vol. LXXI**, (John Wiley & Sons, New York, 1988.)
5. Clarage, J.B., Romo, T., Andrews, B.K., Pettitt, B.M., and Phillips, G.N. *Proc. Natl. Acad. Sci. (USA),* **92**, 3288-3292 (1995).
6. Elber, R. and Karplus, M., *Science,* **235**, 318 (1987)
7. Frauenfelder, H., Siglar,H.A., and Wolynes, P., *Science,* **254**, 1598 (1991)
8. Frauenfelder, H., Steinbach, P.J., and Young, R.D., *Chemica Scripta,* **29A**, 145 (1989).

9. García, A.E. *Phys. Rev. Lett.*, **68**, 2696 (1992).

10. García, A.E. & Stiller, L. *J. Comp. Chem.*, **12**, 1 (1993).

11. García, A.E., Soumpasis, D.M. and Jovin, T.M., *Biophys. J.*, **66**, 1742-1755 (1994).

12. García, A.E. In *Nonlinear Excitations in Biomolecules*, M. Peyrard, Editor. (Springer-Verlag, Berlin, 1995). pp. 191-208.

13. Go, N. and Noguti, T., *Chemica Scripta,* **29A**, 151 (1989).

14. Hendrickson, W.A. and M.M. Teeter, *Nature,* **290**, 107 (1981).

15. Iben, I.E.T. et al., *Phys. Rev. Lett.*, **62**, 1916 (1989).

16. Karplus, M. and Kushick, J.N.,*Macromolecules,* **14**, 325 (1981).

17. Krumhansl, J.A., in *Computer Analysis for Life Science*, edited by C. Kawabata and A.R. Bishop (Ohmsha LTD, Tokyo, Japan, 1985), pp. 78-88.

18. Levart, L., Morineau, A., and Warwick, K. M., *Multivariate Descriptive Statistical Analysis*, (John Wiley & Sons, New York, 1984).

19. Levitt, M., C. Sander, and P.S. Stern, *J. Mol. Biol.,* **181**, 423 (1985).

20. Llinás,M., A. de Marco, and Lecomte, J.T.L.. *Biochem.,* **19**, 1140 (1980).

21. McLachalan, J., *J. Mol. Biol.,* **128**, 49 (1979).

22. Schulz, G.E. and R.H. Schirmer, *Principles of Protein Structure*, (Springer-Verlag, NY, 1978.)

23. Subramaniam, V., Bergenhem, N.C.S., Gafni, A. and Steel, D., *Biochem.* **34**, 1133-1136 (1995).

24. Teeter, M.M., *Proc. Natl. Acad. Sci. (USA),* **81**, 6014 (1984).

25. Teeter, M.M. *Ann. Rev. Biophys. Biophys. Chem.,* **20**, 577 (1991).

26. Teeter, M.M. and D.A. Case, *J. Phys. Chem.,* **94**, 8091 (1990).

27. Usha, M.G. and R.J. Wittebort, *J. Mol. Biol.*, **208**, 669 (1989).

28. Vermeulen,J.A.W.H., R.M.J.N. Lamerichs, 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).

29. Whitlow, M. and M.M. Teeter, *J. Amer. Chem. Soc.,* **108**, 7163 (1986).

5 Figure Captions

Figure 1. Contour plot of the *root-mean-square* distance between pairs of conformations adopted by the protein at 6 ps intervals along the 1200 ps 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 3.00 Å.

Figure 2. Radial tree representation of structures in different clusters. The numbers around the tree show the time (in ps) at which the structure occurs in the trajectory.

Figure 3. Projection $p_i(t)$ of the 1200 ps molecular dynamics trajectory along the five principal MODC 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 Å^2).

Figure 4. Projection of the molecular dynamics trajectory on the plane spanned by directions (a) $\vec{m}_1$ and $\vec{m}_2$, and (b) $\vec{m}_2$ and $\vec{m}_3$, for the first 310 ps of the trajectory.

Figure 5. Projection of the molecular dynamics trajectory on the plane spanned by MODC $\vec{m}_1$ and $\vec{m}_2$, for the first 750 ps trajectory; and (b) $\vec{m}_1$ and $\vec{m}_2$ for the 1.2 ns trajectory.

Figure 6. Time dependence of the velocity autocorrelation function for the five principal MODC.

Figure 7. Log-log plot of the mean square displacements along the five principal MODC as a function of time.

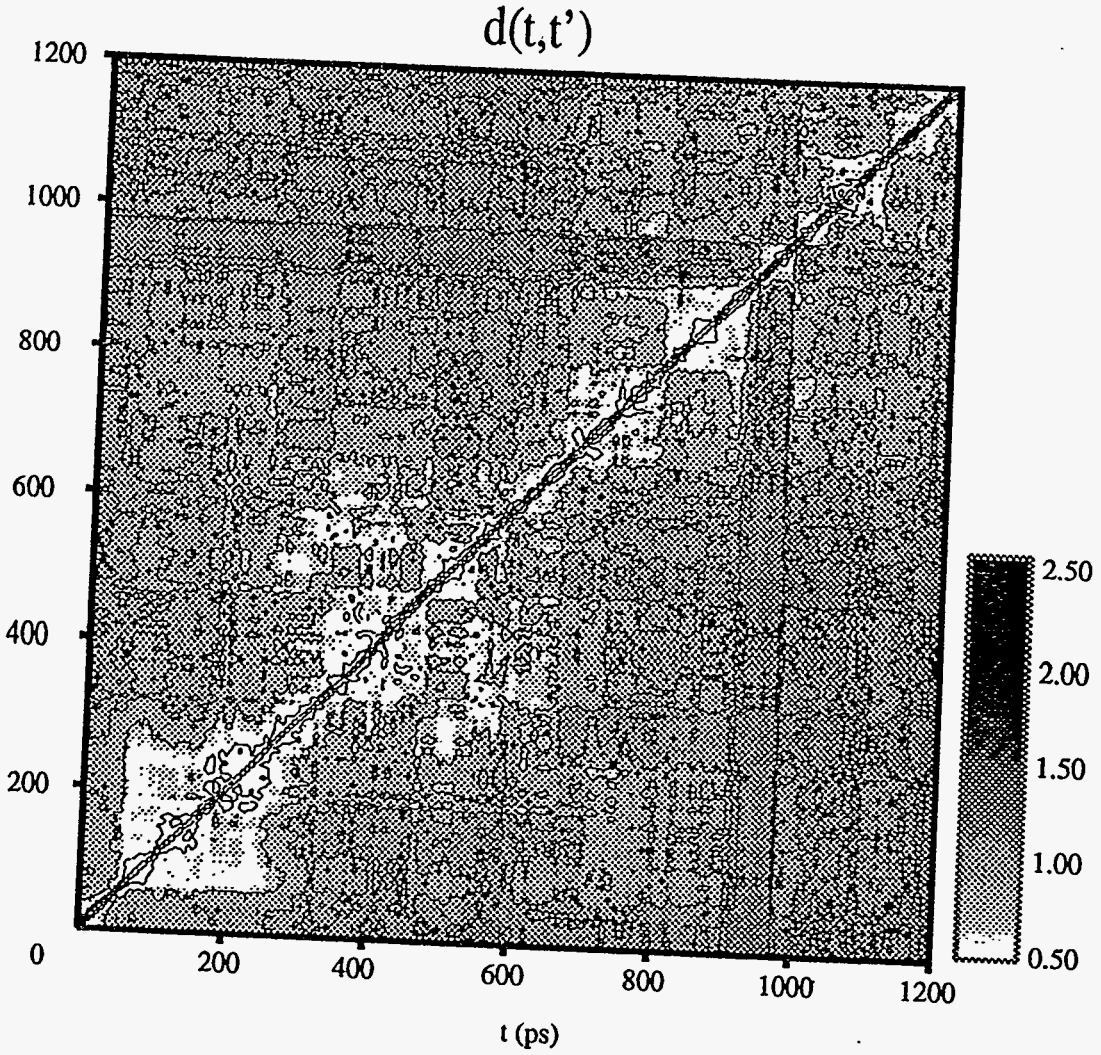

d(t,t')
1200
1000
800
600
400
200
0
200
400
600
800
1000
1200
t (ps)
2.50
2.00
1.50
1.00
0.50

Fig. 1

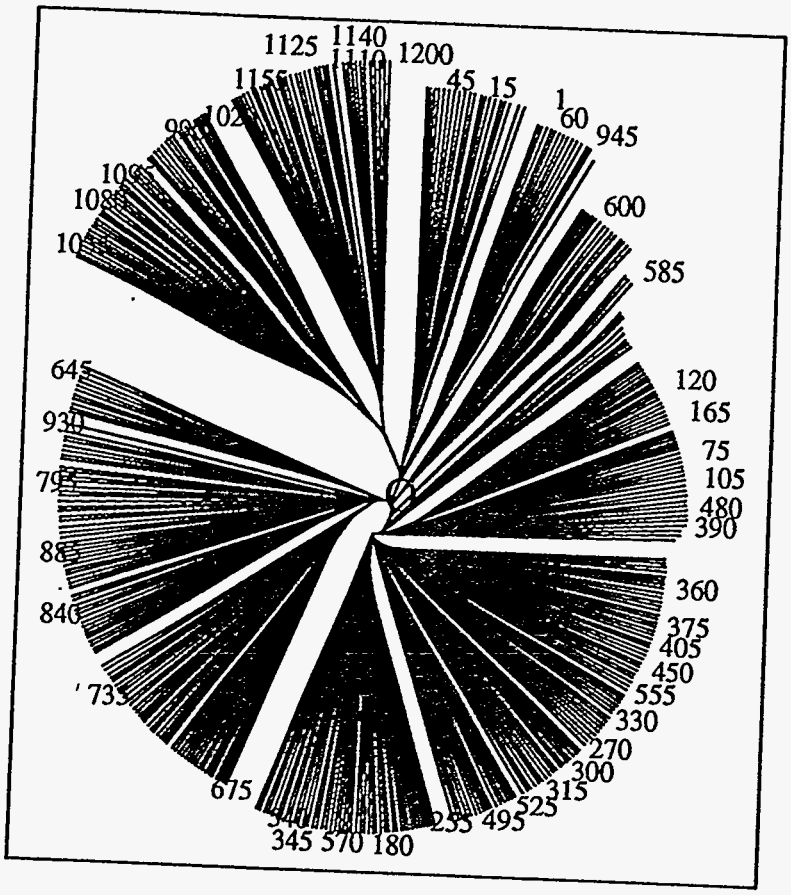

1125
1140
1200
45
15
1
60
945
600
585
120
165
75
105
480
390
360
375
405
450
555
330
270
300
315
525
495
255
180
570
345
675
840
645
930

Fig. 2

CRAMBIN

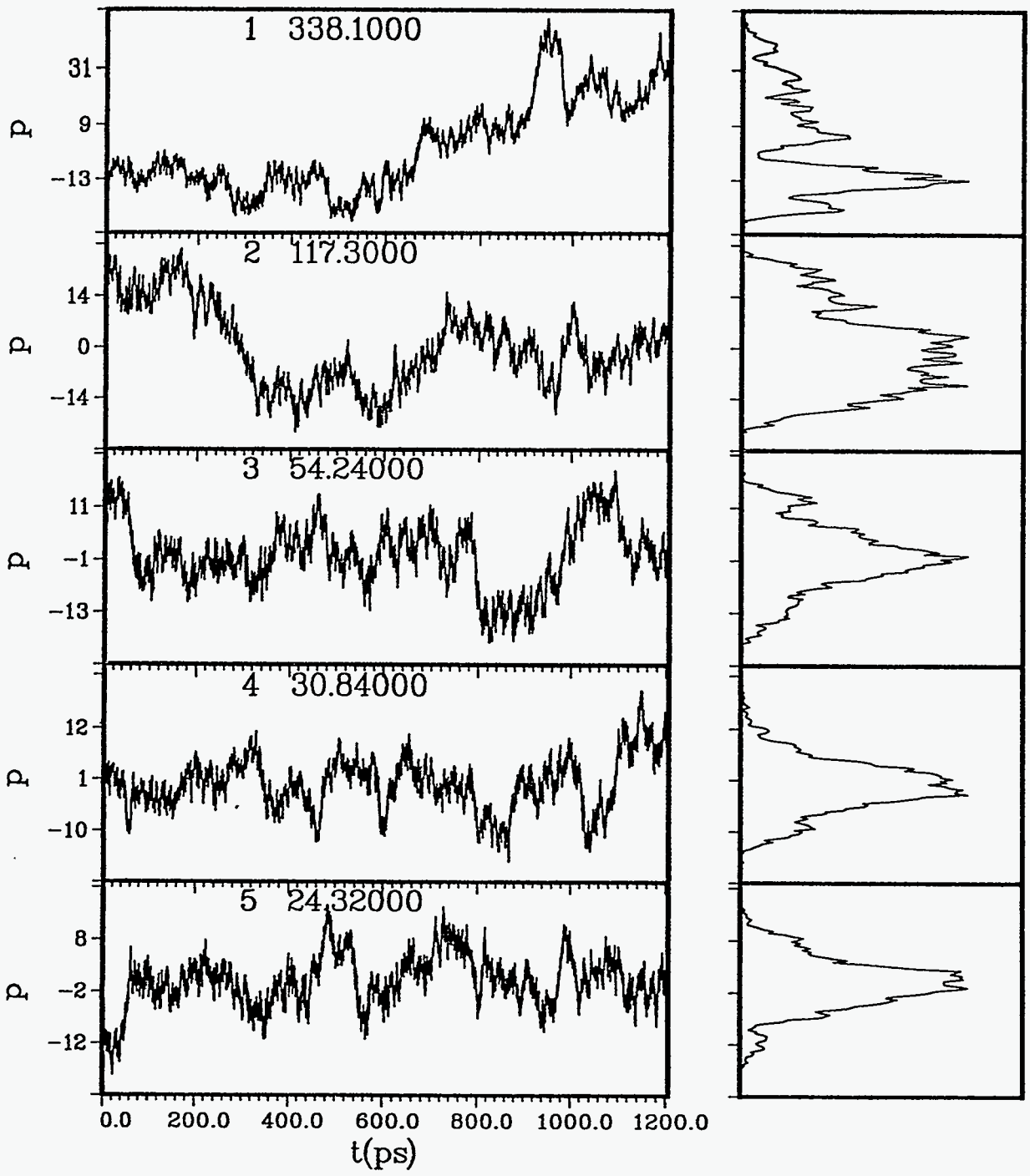
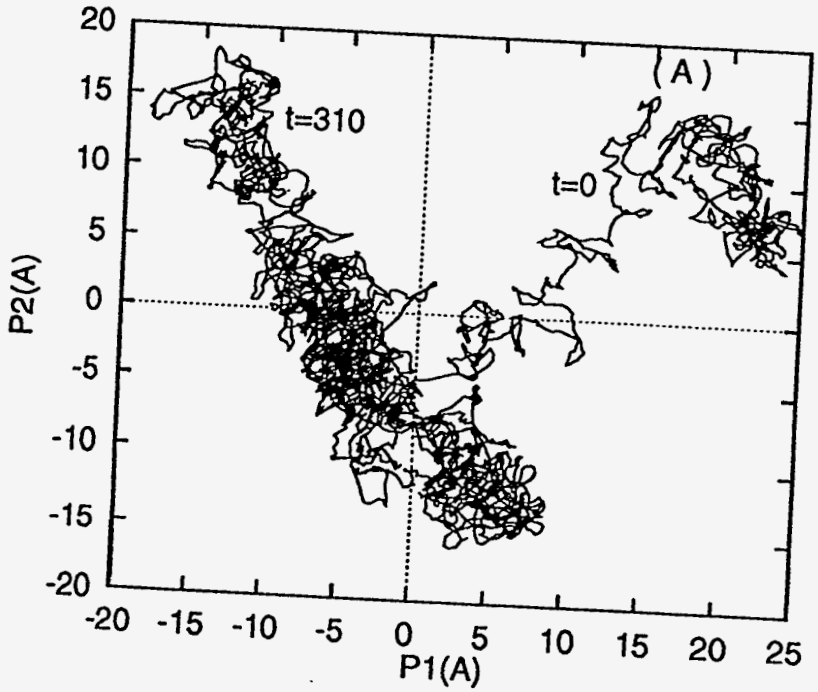


Fig. 3


(A)
t=310
t=0
P2(A)
P1(A)
20
15
10
5
0
-5
-10
-15
-20
-20 -15 -10 -5 0 5 10 15 20 25

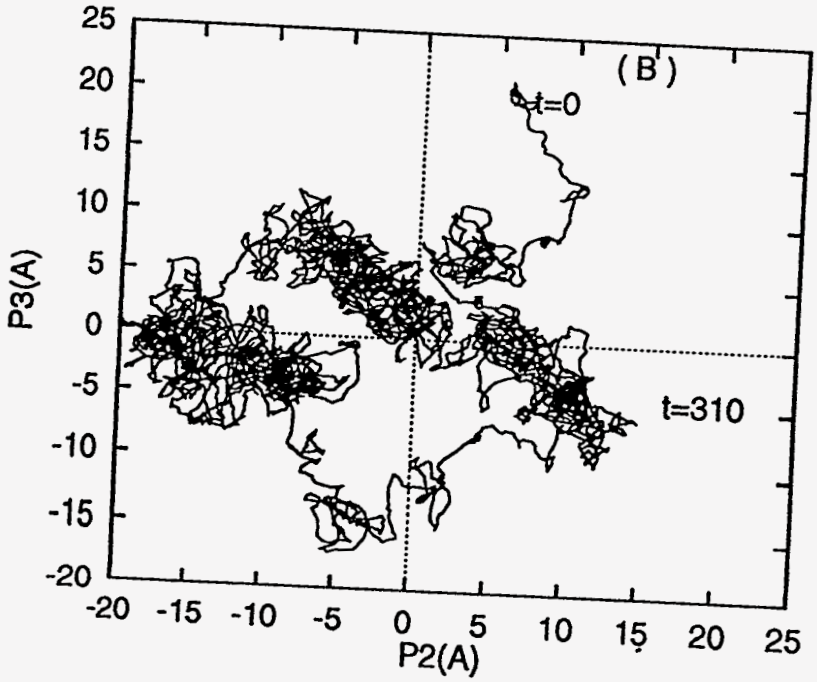
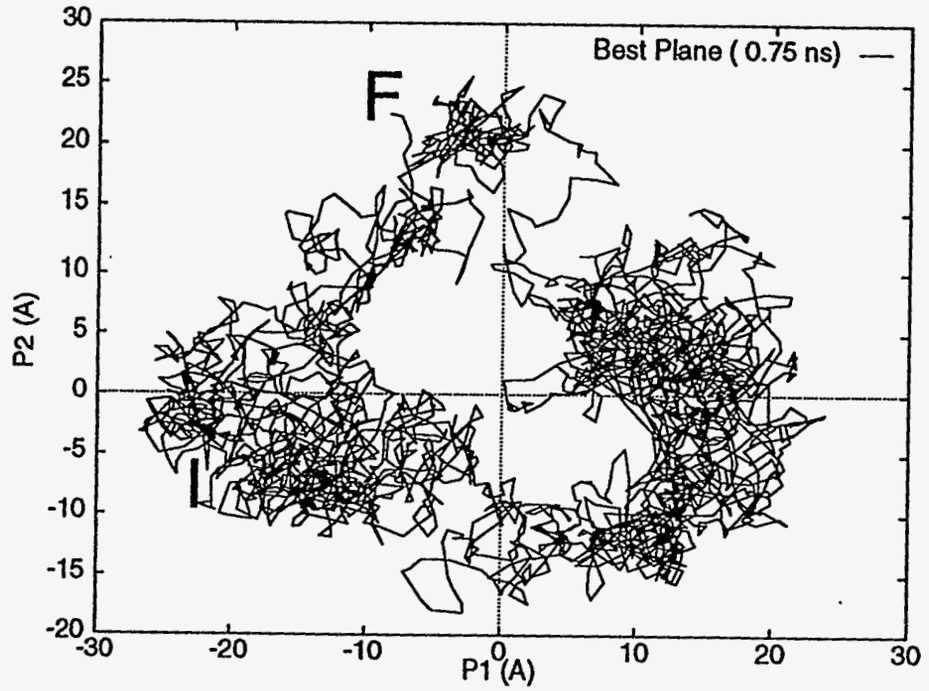

(B)
t=0
t=310
P3(A)
P2(A)
25
20
15
10
5
0
-5
-10
-15
-20
-20 -15 -10 -5 0 5 10 15 20 25

Fig. 4


Best Plane (0.75 ns)
F
I
P2 (A)
P1 (A)
30
25
20
15
10
5
0
-5
-10
-15
-20
-30
-20
-10
0
10
20
30

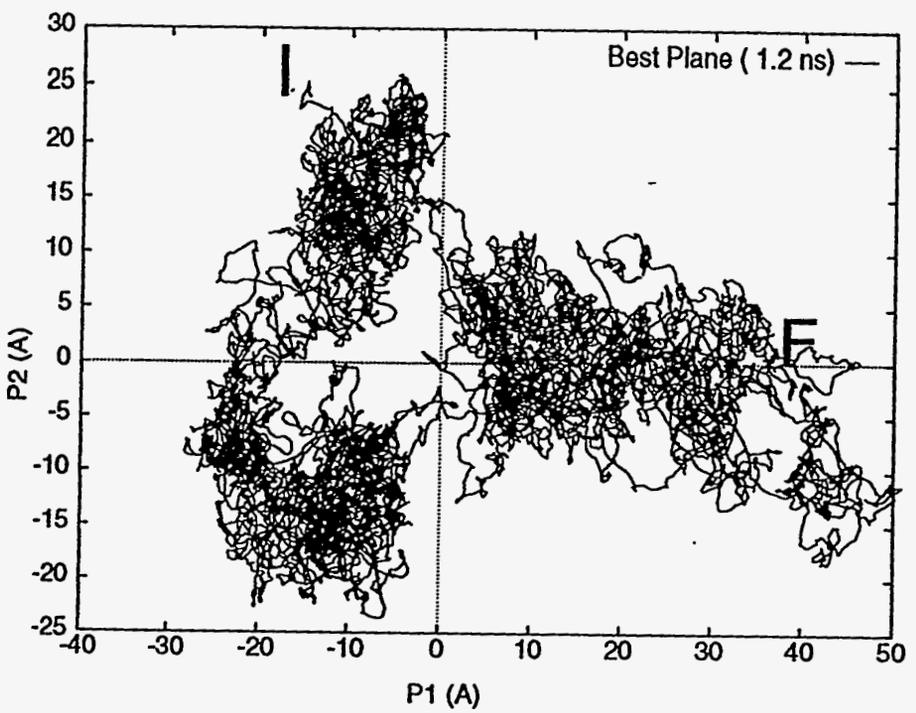

Best Plane (1.2 ns)
I
F
P2 (A)
P1 (A)
30
25
20
15
10
5
0
-5
-10
-15
-20
-25
-40
-30
-20
-10
0
10
20
30
40
50

fig. 5

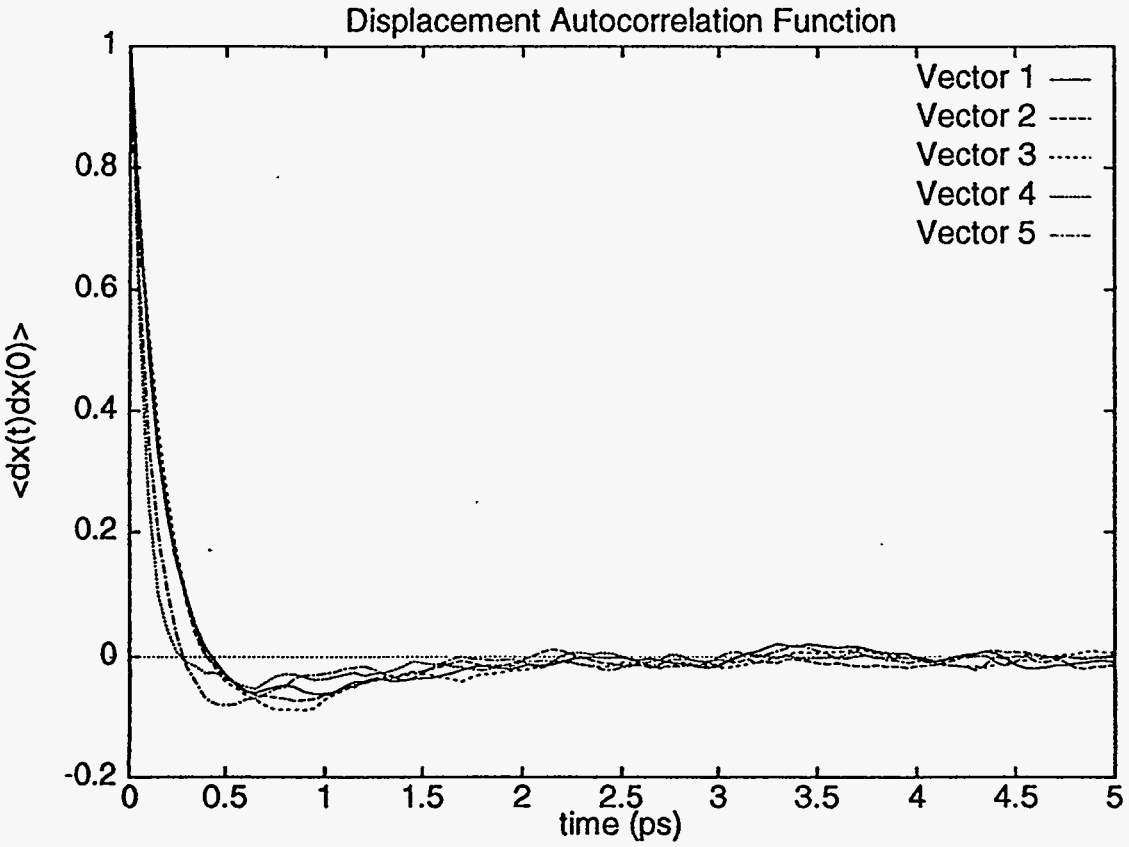

Displacement Autocorrelation Function
Vector 1
Vector 2
Vector 3
Vector 4
Vector 5
<dx(t)dx(0)>
time (ps)
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

fig. 6

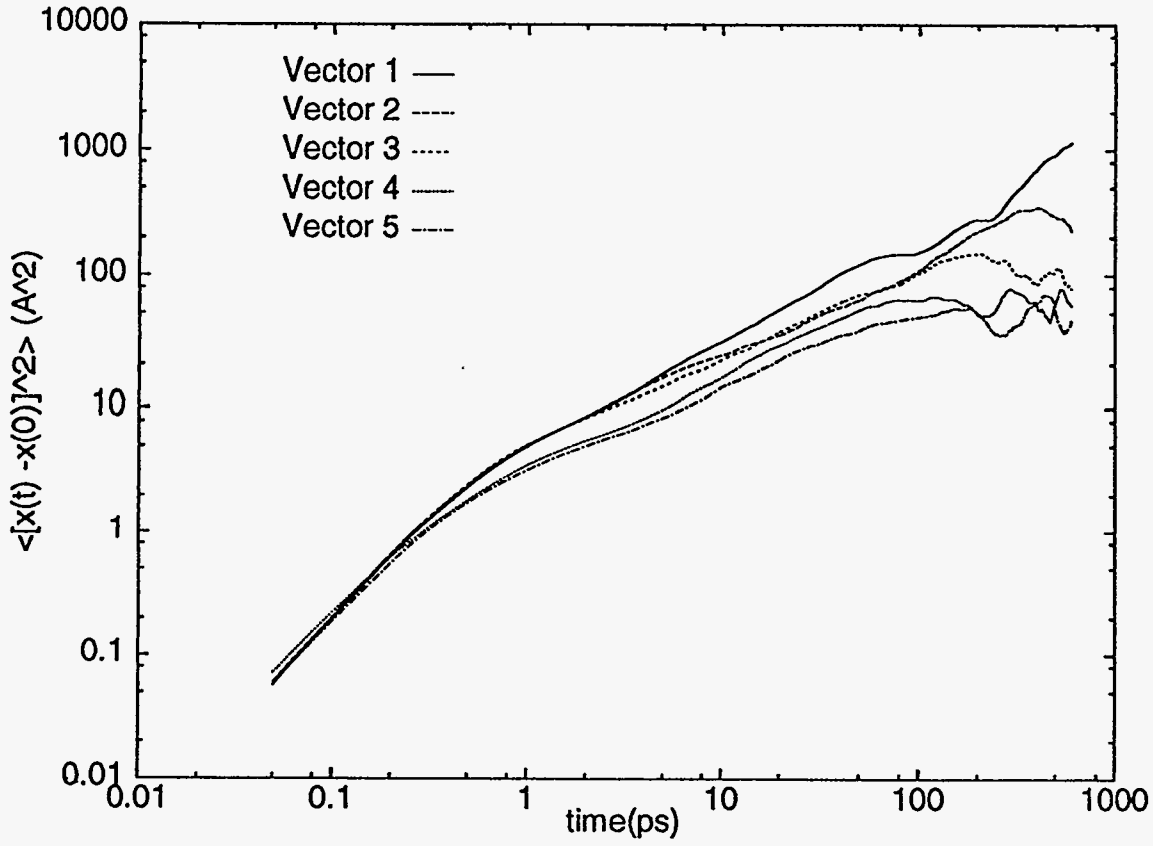

Vector 1
Vector 2
Vector 3
Vector 4
Vector 5
<[x(t) -x(0)]^2> (A^2)
time(ps)
10000
1000
100
10
1
0.1
0.01
0.01
0.1
1
10
100
1000

fig. 7