PROTEINS: Structure, Function, and Genetics 33:496–517 (1998)

# Energy Landscape of a Native Protein: Jumping-Among-Minima Model

**Akio Kitao, Steven Hayward, and Nobuhiro Go***
*Department of Chemistry, Graduate School of Science, Kyoto University, Kyoto, Japan*

***ABSTRACT*** **We have investigated energy landscape of human lysozyme in its native state by using principal component analysis and a model, jumping-among-minima (JAM) model. These analyses are applied to 1 nsec molecular dynamics trajectory of the protein in water. An assumption embodied in the JAM model allows us to divide protein motions into intra-substate and inter-substate motions. By examining intra-substate motions, it is shown that energy surfaces of individual conformational substates are nearly harmonic and mutually similar. As a result of principal component analysis and JAM model analysis, protein motions are shown to consist of three types of collective modes, multiply hierarchical modes, singly hierarchical modes, and harmonic modes. Multiply hierarchical modes, the number of which accounts only for 0.5% of all modes, dominate contributions to total mean-square atomic fluctuation. Inter-substate motions are observed only in a small-dimensional subspace spanned by the axes of multiplyhierarchical and singly hierarchical modes. Inter-substate motions have two notable time components: faster component seen within 200 psec and slower component. The former involves transitions among the conformational substates of the low-level hierarchy, whereas the latter involves transitions of the higher level substates observed along the first four multiply hierarchical modes. We also discuss dependence of the subspace, which contains conformational substates, on time duration of simulation. Proteins 33:496–517, 1998.**


**Key words: energy landscape; hierarchical conformational substates; molecular dynamics; normal mode analysis; principal component analysis; jumping-among-minima model; human lysozyme**

## INTRODUCTION

Protein dynamics involves motions of a great variety of temporal and spatial scales. This originates from the complex nature of the protein energy landscape. For the understanding of protein dynamics in the native state, it is essential to investigate its energy landscape. In this paper, we study the nature of the energy landscape of the native-state protein by using molecular dynamics (MD) simulation.

To understand the nature of the protein energy landscape, we ask the following four questions. Our purpose in this paper is to answer these four questions. We will also discuss the generality of the results.

*Question (1):* Are "conformational substates" mutually similar?

Because of the extreme complexity of protein three-dimensional structures, a large number of substates, which are energetically comparable with one another and distinguishable by their conformations, exist on the protein energy surface in the native state. These substates are commonly termed "conformational substates." To answer question (1), we define conformational substates by introducing physical assumptions. We employ two different models, which will be described in detail in Materials and Methods. In the first model, it is assumed that local energy landscape of each conformational substate is identical with one another. By examining the applicability of the model, we can assess the validity of this assumption. In the second model, we introduce an explicit assumption as to positions of various conformational substates in the conformational space. Then, we examine the local energy landscape of each conformational substate and we assess their mutual similarities.

*Question (2):* How are conformational substates distributed in the multi-dimensional conformational space?

Proteins have a large number of degrees of freedom. In the space spanned by the protein degrees of freedom, are conformational substates distributed in a large-dimensional or in a small-dimensional sub-

*Abbreviations:* MD, molecular dynamics; JAM, jumping-among-minima; SSBP, spherical solvent boundary potential; CMM, cell multipole method; MSF, mean-square fluctuation; RMSF, root-mean-square fluctuation.

Grant sponsor: Ministry of Education, Science and Culture, Japan.
Steven Hayward's present address is Department of Biophysical Chemistry, University of Groningen, The Netherlands.
*Correspondence to: Professor Nobuhiro Go, Department of Chemistry, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan. E-mail: go@qchem.kuchem.kyoto-u.ac.jp
Received 2 March 1998; Accepted 25 June 1998

space? When conformational substates are identified in the conformational space, inter-substate motions can also be defined. From the analysis of inter-substate motions, we can determine the subspace in which conformational substates are distributed. Therefore, by using the two models introduced above, we can answer question (2).

*Question (3):* Are conformational substates hierarchical?

Recently, a hierarchical picture of conformational substates has been inferred from both experimental[1–3] and theoretical[4–10] studies of the native-state protein dynamics. This picture has succeeded in interpreting various phenomena found in a wide temperature range. In the analyses we employ in this paper, we do not make any a priori assumption as to the hierarchical nature of conformational substates. From the results of the analyses, we assess the hierarchical nature of the conformational substates.

*Question (4):* How does the subspace, which contains conformational substates, evolves as a function of duration of simulation time?

This is concerned with practical problems of MD. Since MD length is limited, we should clarify how our results depend on it. This can be done by carrying out the analyses for various MD lengths.

In the analyses to answer the above four questions, we employ two important ideas. One is the collective mode description of protein dynamics and another is the division of protein motions into intra- and inter-substate motions. Both ideas involve a use of a variance-covariance matrix, the matrix of the second moments of atomic coordinates. One main motivation for using this matrix is to introduce a set of collective coordinates, which span a normalized orthogonal system. Principal component analysis is carried out by solving the eigenvalue problem of the second-moment matrix. Projections of atomic coordinates onto collective coordinates, whose axes are defined as eigenvectors of the second-moment matrix, are performed without any loss of information. Since protein dynamics is highly anisotropic, collective mode descriptions of protein motions have been successful in the studies of the protein dynamics.[11–17]

The use of the second-moment matrix has another advantage. This matrix can be divided into two matrices, the matrix of the second moments originated from intra-substate motion, and that originated from inter-substate jumps. In the level of the second moments, cross terms of intra- and inter-substate motions do not appear. We call this new treatment of the second-moment matrix, Jumping-Among-Minima (JAM) model. In Materials and Methods, we will show that second-moment matrix is mathematically divisible into the two terms. Although this comes from the idea of dividing protein motions into intra- and inter-substate motions, the formulation is quite general. In actual application, we introduce a few more physical assumptions. By doing this, we can get a specific physical picture regarding intra- and inter-substate motions.

## MATERIALS AND METHODS

### Simulation

We study the conformational dynamics of human lysozyme in this paper. In this protein and other lysozymes, domain motion, called "hinge-bending motion," has been observed both experimentally[18,19] and theoretically.[17,20–29] It has been shown that this domain motion involves a small energy change and is likely to be responsible for its enzymatic function. We are particularly interested in the hinge-bending motion, because strong anharmonic behaviors including inter-substate motions are expected to occur in a low-energy conformational subspace.

In order to carry out quantitative analyses, we performed MD simulation employing a set of algorithms which have been chosen very carefully by using a newly developed program package[30] base on the framework of Minimization/MD program PRESTO.[70] Spherical Solvent Boundary Potential (SSBP),[31] Cell Multipole Method (CMM),[32] and Nosé-Hoover algorithm[33,34] are employed. High performance of the computer program achieved by the implementation of these algorithms is described elsewhere.[30]

The crystal structure of human lysozyme determined by the method of the normal mode refinement[18] is employed as an initial coordinate of the simulation. Two chloride ions and 146 water molecules (crystal water molecules) are initially placed at the positions determined by the refinement. Then, additional water molecules are placed around human lysozyme to fill a sphere of 34 Å radius. Those water molecules cover lysozyme with at least four layers of water molecules. The total number of explicit water molecules included in the system is 4,684. AMBER potential energy function[35] and TIP3P water model[36] are employed.

MD simulation is performed by the following procedure: (1) Energy minimization of 2,000 steps is carried out with a large value of constraint for restricting lysozyme heavy atoms and crystal water molecules to the initial positions; (2) 50 psec MD is performed in order to equilibrate the system to 300 K and 1 atm, by gradually relaxing the constraints; (3) 50 psec MD is performed without constraints; (4) 1 nsec MD is carried out. The numerical integration of the equation of motion is carried out by Gear predictor-corrector method[37] with a time step of 0.5 fsec. The coordinate and velocity trajectory are stored at every 5 fsec. They are used for the following analyses. Therefore, number of stored instantaneous conformations, $K$, is $2 \times 10^5$.

Normal mode analysis in vacuum is also carried out in the following procedure: (1) Energy minimization is performed starting from the average struc-

ture of 1 nsec MD; (2) Normal modes are calculated for the minimum energy conformation by using a standard procedure.[38] Normal modes, which are determined only by the intra-molecular interactions within a protein, are employed in the following analyses as a specific coordinate system to understand behavior of a complicated real system.

### Treatment of Protein Degrees of Freedom

In this section, we describe the treatment of protein degrees of freedom in this paper. In the following, only protein internal degrees of freedom are extracted from the MD trajectory and are used for further analysis. An instantaneous conformation at time $t$ is described in terms of mass-weighted position vectors, $\mathbf{x}_j(t)$, $j = 1, 2, \ldots, N$, or mass-weighted Cartesian coordinates of constituent atoms, $x_i(t)$, $i = 1, 2, \ldots, 3N$, where $N$ is the number of atoms in a protein, $\mathbf{x}_j(t)$ is position vector of the $j$th atom times square-root of (its mass over total mass of protein), and $x_{3j-2}(t)$, $x_{3j-1}(t)$, and $x_{3j}(t)$ are x-, y-, and z-component of $\mathbf{x}_j(t)$, respectively. External degrees of freedom are removed first by placing the center of gravity of the molecule always at the origin of the coordinate system and further by imposing the Eckart condition,[39]

$$\sum_j \mathbf{x}_j \times \mathbf{x}_j^0 = 0, \tag{1}$$

where $\mathbf{x}_j^0$ is the value of $\mathbf{x}_j$ at its reference state, which is taken from a representative minimum energy conformation, representative in the sense to be as similar as possible to the time-average of $\mathbf{x}_j$.

### Second-Moment Matrix and Principal Component Analysis

Here we briefly describe principal component analysis. Principal component analysis is a powerful method to analyze a region of conformational space explored by a trajectory of state point.[12,14] Principal component is determined as a solution of eigenvalue problem of the second-moment matrix. Therefore we will study the following second-moment matrix, **A**, whose matrix element, $a_{ij}$, is,

$$a_{ij} \equiv \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle, \tag{2}$$

where $\langle \cdots \rangle$ means average over $K$ instantaneous structures sampled during the period of simulation. If the energy surface were a multi-dimensional parabola and sampling is sufficiently done, the second-moment matrix should be given by,

$$a_{ij} = \sum_l w_{il}^N w_{jl}^N \lambda_l^N, \tag{3}$$

where $\lambda_l^N$ is the mean-square-fluctuation (MSF) of the $l$th normal mode, which is given by $\lambda_l^N = k_B T/\omega_l^2$. Here $\omega_l$ is the angular frequency of the $l$th normal mode, $k_B$ is the Boltzmann constant, $T$ is the absolute temperature, and $w_{il}^N$ is the $i$th element of the $l$th normal mode (N) eigenvector.

### Jumping-Among-Minima (JAM) Model

Here we introduce jumping-among-minima (JAM) model, which is the main model employed in the following analyses. In this section, we show that the second-moment matrix is divisible into two terms, the matrix of second moments determined by inter-substate motions and that determined by intra-substate motions. Although we have this physical picture in our mind, the formulation itself is quite general and applicable to many cases.

At first we divide the whole native-state conformational space into catchment regions, each of which contains one energy minimum. According to this division, $K$ instantaneous conformations are classified into $M$ groups. Instantaneous conformations belonging to the same catchment region are classified into the same group. Then Equation 2 is written as,

$$a_{ij} = \sum_{k=1}^{M} f_k \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle_k. \tag{4}$$

Let $n_k$ be the number of conformations in the $k$th group, the average $\langle \cdots \rangle_k$ is taken over $n_k$ conformations in the $k$th group, and,

$$K = \sum_{k=1}^{M} n_k, \; f_k = \frac{n_k}{K}. \tag{5}$$

Then the right-hand-side of Equation 4 can be divided into two terms,

$$a_{ij} = \sum_{k=1}^{M} f_k (\langle x_i \rangle_k - \langle x_i \rangle)(\langle x_j \rangle_k - \langle x_j \rangle) + \sum_{k=1}^{M} f_k \langle (x_i - \langle x_j \rangle_k)(x_j - \langle x_j \rangle_k) \rangle_k. \tag{6}$$

The first term of Equation 6 represents contribution of atomic fluctuations caused by jumping-among-minima. The second term represents contribution of atomic fluctuations within each catchment region. This model can be further extended to a hierarchical model by repeatedly applying similar operation used in the derivation of Equation 6 to the first term of the right-hand-side of Equation 6.

Up to this point, no physical assumption is made as to the nature of the catchment regions. The only thing employed in the derivation of Equation 6 is that instantaneous conformations are classified into groups. In actual application to the analysis of protein energy surface, we introduce strong physical

assumptions in order to obtain a physical picture. We examine two methods of applying JAM model to real system as will be described in the following two sections.

### JAM(I) Model: An Idealized Case Where Local Energy Landscape of Each Catchment Region is Identical and Harmonic

Let us consider a case, where local energy landscape of each catchment region is mutually very similar and is harmonic. In this idealized case of identical local landscape, Equation 6 reduces to,

$$a_{ij} = \sum_k f_k (x_i^k - \langle x_i \rangle)(x_j^k - \langle x_j \rangle) + \sum_l w_{il}^N w_{jl}^N \frac{k_B T}{\omega_l^2}, \quad (7)$$

where $x_i^k$ is the $i$th coordinate of the $k$th minimum. Derivation of Equation 7 is described in Appendix A. This equation is derived based on the assumptions shown by Equations A3 and A6. These assumptions are valid if the state point stays in each catchment region for a significant period longer than the characteristic time of intra-substate motions. The contribution to second moment from the JAM motion, $\mathbf{S}^I = (s_{ij}^I)$, is defined as,

$$\begin{aligned} s_{ij}^I &= \sum_k f_k (x_i^k - \langle x_i \rangle)(x_j^k - \langle x_j \rangle) \\ &= a_{ij} - \sum_l w_{il}^N w_{jl}^N \frac{k_B T}{\omega_l^2}, \end{aligned} \quad (8)$$

where the first term, $a_{ij}$, is directly determined by MD and the second term is obtained by normal mode analysis. It should be noted that an explicit expression of the matrix $\mathbf{S}^I$ is determined without identifying the positions of catchment regions. This was made possible by the introduction of strong physical assumptions about the local energy landscape. JAM model of the idealized case described here is termed JAM(I) in the following sections.

### JAM(R) Model: Rotamer States as Conformational Substates

This model is defined by introducing an explicit assumption on boundaries between catchment regions in the native-state conformational space. For this purpose, values of all rotatable dihedral angles in the protein are monitored during the 1 nsec MD simulation. Range of variation of each dihedral angle is divided into a few rotamer states, details of which will be described later in a section in Results. Then, the multi-dimensional conformational space of the protein is divided into subspaces, each of which is characterized by a set of rotamer states of individual dihedral angles. We call each of such subspaces a rotamer state of protein as a whole, or a protein rotamer state. Now we introduce an assumption that a protein rotamer state is a good approximation of a catchment region. In other words, we assume that a conformational state point stays within a catchment region, unless any one of rotatable dihedral angles crosses over a boundary between rotamer states. For this explicit assumption about catchment regions, Equation 6 is rewritten as,

$$a_{ij} = s_{ij}^R + \sum_{k=1}^M f_k r_{ij}^k, \quad (9)$$

where $s_{ij}^R$ is a contribution from transitions between protein rotamer states and $r_{ij}^k$ is a contribution from the $k$th protein rotamer state. Both matrices $\mathbf{S}^R = (s_{ij}^R)$ and $\mathbf{R}^k = (r_{ij}^k)$ are determined directly from MD trajectory. This JAM model is termed JAM(R) in the following.

### Principal Mode and JAM Mode

From the matrices of the second moments described above, we introduce collective coordinates. The matrix of principal component (P) eigenvectors, $\mathbf{W}^P$, is determined as a solution of the standard eigenvalue problem,

$$\mathbf{A}\mathbf{W}^P = \mathbf{W}^P \zeta^P, \quad (10)$$

where $\zeta^P = \mathrm{diag}(\zeta_m^P)$ is a diagonal matrix whose diagonal element, $\zeta_m^P$, is the $m$th eigenvalue. Eigenvalues and corresponding eigenvectors are numbered according to the magnitude of eigenvalues in descending order. Protein motion along the $m$th eigenvector is termed the $m$th principal mode. The quantity $\zeta_m^P$ represents the MSF of the $m$th principal mode. Similarly, eigenvectors and eigenvalues of JAM(I) mode, JAM(R) mode, and principal mode of the $k$th rotamer state are determined by eigenvalue problems,

$$\mathbf{S}^I \mathbf{W}^I = \mathbf{W}^I \lambda^I, \quad (11)$$

$$\mathbf{S}^R \mathbf{W}^R = \mathbf{W}^R \lambda^R. \quad (12)$$

$$\mathbf{R}^k \mathbf{W}^k = \mathbf{W}^k \lambda^k. \quad (13)$$

The eigenvalues $\lambda_m^I$ and $\lambda_m^R$, represent MSF along the $m$th JAM(I) and JAM(R) modes originated only from the JAM(I) and JAM(R) motions, respectively. The quantity $\lambda_m^k$ represents MSF along the $m$th principal mode of the $k$th rotamer state. Eigenvectors satisfy the following orthonormal condition,

$$(\mathbf{W}^\alpha)^t \mathbf{W}^\alpha = \mathbf{I}, \quad (14)$$

where $\mathbf{I}$ is a unit matrix, $\mathbf{W}^\alpha$ is a matrix of eigenvectors, $\mathbf{W}^N$, $\mathbf{W}^P$, $\mathbf{W}^I$, $\mathbf{W}^R$, or $\mathbf{W}^k$ in which $N$, $P$, $I$, $R$, and $k$ stand for normal mode, principal mode, JAM(I)

mode, JAM(R) mode and principal mode of the $k$th rotamer state, respectively.

In addition to $\zeta_m^P$, we also define the quantities $\zeta_m^N$, $\zeta_m^I$, and $\zeta_m^R$, which represent the $m$th diagonal elements of the matrices, $(\mathbf{W}^N)^t\ \mathbf{A}\mathbf{W}^N$, $(\mathbf{W}^I)^t\ \mathbf{A}\mathbf{W}^I$, and $(\mathbf{W}^R)^t\ \mathbf{A}\mathbf{W}^R$, respectively. These quantities $\zeta_m^N$, $\zeta_m^I$, and $\zeta_m^R$, are interpreted as total (JAM + intra-substate) MSF along the $m$th mode axes of the normal mode, JAM(I) mode, and JAM(R), respectively. As in the case of $\mathbf{W}^\alpha$, we use similar expressions, $\lambda_m^\alpha$ and $\zeta_m^\alpha$.

Four different collective coordinate sets, i.e., normal mode, principal mode, JAM(I) mode, and JAM(R) mode coordinate sets, are used in this paper. To understand relations among these coordinates, similarity of two coordinate axes is quantified by their inner products.[38,40,41] For example, the $m$th mode vector, $\mathbf{w}_m^\alpha$ of the coordinate set $\alpha$ ($\alpha$ is $N$, $P$, $I$, or $R$) is expressed as a linear combination of the $n$th eigenvector of the coordinate set $\beta$, $\mathbf{w}_n^\beta$,

$$\mathbf{w}_m^\alpha = \sum_{n=1}^{3N-6} g_{mn}^{\alpha\beta}\mathbf{w}_n^\beta = \sum_{n=1}^{3N-6} (\mathbf{w}_m^\alpha, \mathbf{w}_n^\beta)\mathbf{w}_n^\beta. \quad (15)$$

where $(\mathbf{w}_m^\alpha, \mathbf{w}_n^\beta)$ represents an inner product of $\mathbf{w}_m^\alpha$ and $\mathbf{w}_n^\beta$. Here we introduce a coefficient,

$$c_{mn}^{\alpha\beta} = (g_{mn}^{\alpha\beta})^2. \quad (16)$$

The coefficient $c_{mn}^{\alpha\beta}$, which has a value in the range $1 \geq c_{mn}^{\alpha\beta} \geq 0$, represents the magnitude of contribution of the vector $\mathbf{w}_n^\beta$ to the vector $\mathbf{w}_m^\alpha$. When $c_{mn}^{\alpha\beta}$ is unity, two vectors are identical. To show a relation of a subspace spanned by a certain subset of vectors, $[\mathbf{w}_n^\beta]$, with a vector $\mathbf{w}_m^\alpha$, it is convenient to define the quantity,

$$P_m^{\alpha\beta} = \sum_{n\in[\mathbf{w}\beta/n]} c_{mn}^{\alpha\beta}. \quad (17)$$

When summation of Equation 17 is taken over all $(3N - 6)$ vectors, $P_m^{\alpha\beta}$ becomes unity,

$$P_m^{\alpha\beta} = \sum_{n=1}^{3N-6} c_{mn}^{\alpha\beta} = 1. \quad (18)$$

### Anharmonicity Factor and Probability Distribution Function

In this section, we introduce a few quantities which will be used in Results. "Anharmonicity factor" is a quantity to measure a degree of anharmonic nature of a mode.[16] To define this factor, we at first define harmonic MSF along the $m$th mode axis of the coordinate system $\alpha$ due to harmonic fluctuations, which is given by,

$$\lambda_m^{\alpha,har} = \sum_{n=1}^{3N-6} c_{mn}^{\alpha N}\lambda_n^N. \quad (19)$$

The harmonic MSF $\lambda_m^{\alpha,har}$ is a MSF along the $m$th mode axis of the coordinate system $\alpha$ if energy surface is harmonic as assumed in the normal modes analysis. Anharmonicity factor is defined as,

$$\rho_m^\alpha = (\zeta_m^\alpha/\lambda_m^{\alpha,har})^{1/2}. \quad (20)$$

This quantity is unity when energy surface is harmonic as assumed in normal mode analysis.

Next, we consider the probability distribution functions of the projections onto collective coordinate space. Projection of MD trajectory onto the $m$th mode of the coordinate system$\alpha$ is defined as,

$$q_m^\alpha(t) = ((\mathbf{x}(t) - \langle\mathbf{x}\rangle), \mathbf{w}_m^\alpha). \quad (21)$$

By using a trajectory of the projection, probability distribution function along the $m$th mode of the coordinate system $\alpha$, $P_m^\alpha(q_m^\alpha)$, is calculated. The quantity, $P_m^\alpha(q_m^\alpha)\ dq_m^\alpha$, represents the probability that the projection is at the position $q_m^\alpha$. We define "effective free energy" along the $m$th mode of the coordinate system$\alpha$, $\mu_m^\alpha(q_m^\alpha)$, by,

$$\mu_m^\alpha(q_m^\alpha) = -k_B T \ln P_m^\alpha(q_m^\alpha). \quad (22)$$

This function gives a one-dimensional effective free-energy curve as a function of $q_m^\alpha$. The function $\mu_m^\alpha(q_m^\alpha)$ is employed in order to understand the energy landscape along this axis. If $P_m^\alpha(q_m^\alpha)$ is a gaussian distribution with a variance $\zeta_m^\alpha$, it is given by,

$$P_m^{\alpha,gau}(q_m^\alpha) = (2\pi\zeta_m^\alpha)^{-1/2} \exp[-(q_m^\alpha)^2/(2\zeta_m^\alpha)]. \quad (23)$$

When energy surface is harmonic as expected from the harmonic MSF of Equation 19, i.e., $\rho_m^\alpha = 1$, $P_m^\alpha(q_m^\alpha)$ should be given by,

$$P_m^{\alpha,har}(q_m^\alpha) = (2\pi\lambda_m^{\alpha,har})^{-1/2} \exp[-(q_m^\alpha)^2/2\lambda_m^{\alpha,har})]. \quad (24)$$

## RESULTS

Before going into the details of the results, we briefly describe the construction of the sections in Results.

In Section A, we describe how we identify rotamer states. The rate coefficient for the rotamer transition is also discussed.

In Section B, in order to answer Question 1, we demonstrate the mutual similarity of local energy landscape of each rotamer state defined in Section A by showing time correlation function of principal mode variables, frequency distribution, and atomic mean-square fluctuation.

**TABLE I. Summary of Dihedral Angle Transitions**

| Type | Number of angles | Number of angles in which number of transitions is 0 | 1 | >1 | Average residence time (psec) |
|---|---|---|---|---|---|
| Non-terminal main-chain dihedral angles | | | | | |
| $\phi$ | 127 | 124 | 1 | 2 | — |
| $\psi$ | 129 | 125 | 1 | 3 | — |
| $\omega$ | 129 | 129 | 0 | 0 | — |
| Dihedral angles of side-chain terminals, N-terminus, and C-terminus | | | | | |
| $-OH$ | 17 | 8 | 0 | 9 | 22.6 |
| $-CH_3$ | 63 | 0 | 0 | 63 | 33.1 |
| $-NH_2$ | 44 | 44 | 0 | 0 | — |
| $-NH_3$ | 6 | 0 | 0 | 6 | 19.9 |
| $-CO_2$ | 12 | 8 | 1 | 3 | — |
| Other side-chain dihedral angles average | 350 | 266 | 15 | 69 | 73.8 |

In Section C, principal modes are classified into three types of modes, multiply-hierarchical modes, singly-hierarchical modes, and harmonic modes.

In Section D, relations among normal mode, principal mode, JAM(I) mode, and JAM(R) mode are discussed. By using these relations, we show that conformational substates are distributed in a small-dimensional subspace. This gives the answer to Question 2.

In Section E, we analyze hierarchical structure of the conformational energy landscape. We will give further characterization of the two types of anharmonic modes, multiply-hierarchical modes and singly-hierarchical modes. Thus, the answer to Question 3 will be given in this section.

To answer Question 4, time evolution of JAM(R) space is discussed in Section F. We show that subspace in which inter-substate motions take place does not change significantly in the time range from 200 to 1000 psec. This provides the answer to Question 4.

### A. Analysis of Dihedral Angle Transitions and Identification of Rotamer States

Before going into the collective mode description of protein dynamics, fluctuations of dihedral angles are briefly mentioned in this section. When MD trajectory is expressed as variations of a set of protein internal coordinates (bond lengths, bond angles, and dihedral angles), those of dihedral angles have the most dominant effects for three-dimensional structure of proteins. When we trace variations of dihedral angles in MD trajectory, we see that for most of the time duration of the simulation a typical dihedral angle fluctuates within a limited range of about 5 to 15 degrees, but occasionally it undergoes a transition of as much as 120 or 180 degrees. Thus, range of variation of each dihedral angle is divided into a few rotamer states. Large amplitude transitions between rotamer states very likely involve crossing over an energy barrier. Therefore, it is reasonable to assume that a conformational state point stays within a catchment region, unless any one of rotatable dihedral angles undergoes a transition between rotamer states. Analysis of variation in dihedral angles shown in this section is a basis of the JAM(R) model in the following analyses.

From the trajectory of 877 dihedral angles, distributions of values of dihedral angles are calculated. Since distributions have clear peaks and minima and most of the transitions occur instantly within ~100 fsec, dihedral angle transitions are easily defined from dihedral angle trajectory. In Table I, summary of the dihedral angle transitions is shown. Except for $\psi$-angle transition of N118, transitions of $\psi$ and $\phi$ angles occur concomitantly in pairs, $\psi$ of V74 and $\phi$ of N75, $\psi$ of I89 and $\phi$ of A90, and $\psi$ of V125 and $\phi$ of Q126. These residues are located in relatively flexible regions, at hinge region (I89, A90) and both sides of the C-terminal helix (N118, V125, Q126). Backbone around V74 and N75 is also known to be flexible, where the C = O group of V74 points toward a sodium ion found inside of the protein[42] whereas this group faces outside in the absence of sodium ion.[43] These transitions in main-chain dihedral angles are not related directly to large conformational changes shown in the following sections.

Based on the above analysis, we define rotamer state of lysozyme. A protein rotamer state is defined as a state in which all the dihedral angles are in the same states. When dihedral-angle transition occurs somewhere in lysozyme, it is considered that a rotamer state changes to another. In the assignment of rotamer states, transitions of 142 terminal dihedral angles of side-chains and main-chain terminals in Table I are not considered. During 1 nsec simulation, 1,171 transitions are detected and $2 \times 10^5$ instantaneous conformations are classified into 1,172

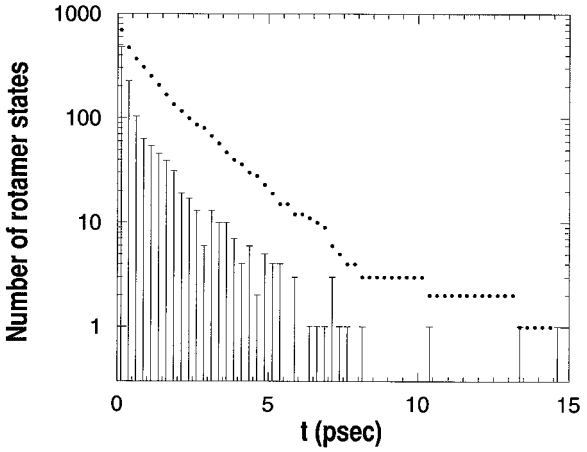


Fig. 1. Histogram of residence time in rotamer states (bars) and decay of the number of rotamer states as a function of time (dots).

protein rotamer states. In Figure 1, distribution of residence time in a rotamer state, $\tau_k$, is shown. Average residence time, $\langle\tau_k\rangle$, and the longest residence time $\tau_{683}$ are 0.85 ± 1.76 and 14.71 psec, respectively. The decay of the number of rotamer states, $[A]_t$, is also shown in Figure 1. $[A]_t$ is given as the number of rotamer states, whose $\tau_k$ is large than $t$. Its time behavior is well expressed by a single exponential. From the single exponential decay, rotamer state transition is understood as a first-order reaction, $d[A]_t/dt = -k[A]_t$. The rate coefficient is, $k = 1/\langle\tau_k\rangle = 1.18 \times 10^{12}$ (sec$^{-1}$). If $k$ for each rotamer state differs significantly from one to another, the decay of $\tau_k$ is not well expressed by a single exponential. From this single exponential decay behavior, it is expected that $k$ for each rotamer state does not differ largely. This suggests that a barrier height from a rotamer state to others does not differ largely from one to another.

### B. Dynamics in Rotamer States

To understand fluctuation within a protein rotamer state, we choose the 683$^{rd}$ rotamer state whose residence time is the longest (14.71 psec). First of all, we carry out principal component analysis of the matrix $\mathbf{R}^{683}$. In Figure 2, examples of autocorrelation functions of projection onto a principal mode are shown. These correlation functions in Figure 2 undergo underdamping. Correlation functions of projections onto all principal mode axes, except for the first two mode axes, are also underdamping.

The behavior of the projection onto principal mode axes can be fairly well reproduced by that of solution of Langevin equation with a harmonic potential. The correlation function for the solution of Langevin equation is given by Equation B6 in Appendix B and in the References 12, 14, 44, and 45. In Figure 2, autocorrelation functions directly calculated from

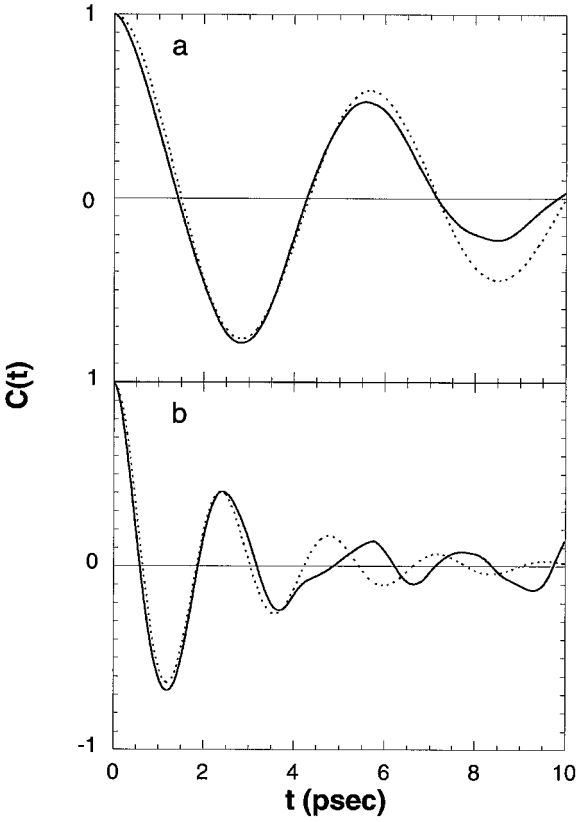


Fig. 2. Autocorrelation functions of projection onto the **(a)** fourth and **(b)** tenth principal modes determined by principal component analysis of the 683rd rotamer state. Those determined from MD trajectory (solid lines) and from Langevin equation (broken lines) are shown. Effective frequency $\omega_0$ and friction coefficient $\gamma$ are 5.9 and 1.0 cm$^{-1}$ for (a), and 14.1 and 4.0 cm$^{-1}$ for (b), respectively.

MD is compared with theoretical correlation functions of Equation B6. In the theoretical correlation functions, angular frequency $\omega_0$ is set to $(k_BT/\lambda_m^{683})^{1/2}$. The quantity, $\lambda_m^{683}$, is determined by Equation 13, which is a MSF along the $m$th principal mode in the 683rd protein rotamer state. The value of friction constant $\gamma$ is chosen so as best to reproduce the correlation function calculated from MD. To reproduce underdamping behaviors as shown in Figure 2, low friction values were used. The good agreements between autocorrelation functions directly calculated from MD and theoretical correlation functions indicate that the local energy surface in the 683rd catchment region is nearly harmonic.

Three different frequency distributions determined from fluctuations in the 683rd, 559th ($\tau_{559}$ = 13.30 psec), and 883rd ($\tau_{883}$ = 10.34 psec) rotamer states, respectively, are shown in Figure 3a. Those from fluctuations in 683rd rotamer state, and in 1 nsec MD, and by normal mode analysis, respectively, are shown in Figure 3b. These distributions are obtained as mode number density. In the case of

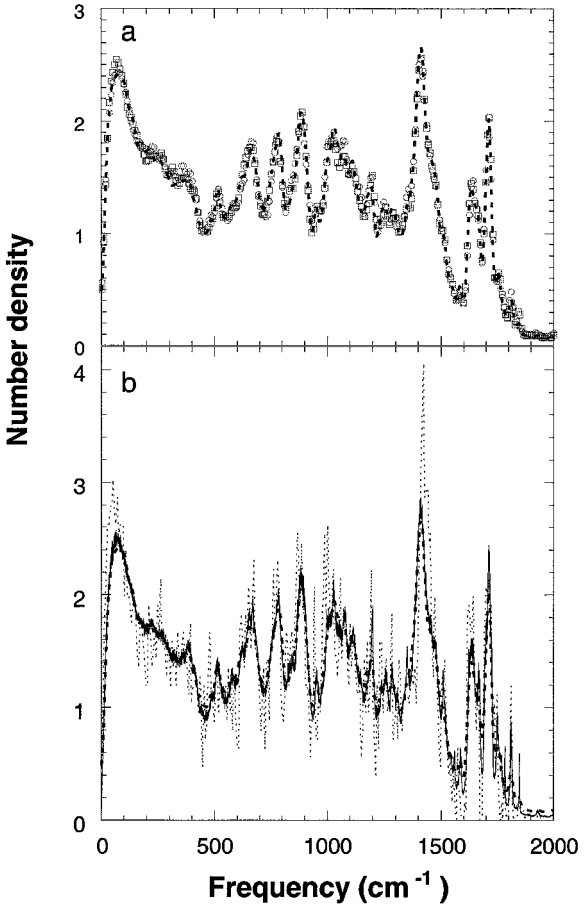


Fig. 3. Mode number densities. **(a)** Those obtained from the 683rd (thick broken line), 559th (□), and 883rd (○) rotamer-state fluctuations. **(b)** Those obtained from the 683rd rotamer-state fluctuation (thick broken line), 1 nsec MD (thin solid line), and normal mode (thin broken line).

one-dimensional Langevin equation with harmonic potential, it is given by Equation B7 in the Appendix B.[12,14] Friction coefficient $\gamma$ significantly affects spectral shape when $\gamma > 2\omega_0$ (see Figure 20 of the Reference 12). Frequency distributions obtained from the three rotamer states are mutually almost identical and also with that obtained from 1 nsec trajectory. This suggests that power spectral densities in the other rotamer states are similar to that of the 683rd rotamer state. From this analysis it is expected that energy surface of each catchment region is mutually similar. As expected from such underdamping behavior as shown in Figure 2 in most (in fact, except for only two) of the modes, mode number densities determined from MD trajectory are in fairly good agreement with normal mode number density. This is a consequence of relatively low friction exerted on protein motion and harmonic nature of protein rotamer states.

Mutual similarity of the local energy landscape is studied also by the analysis of MSF. MSF is very

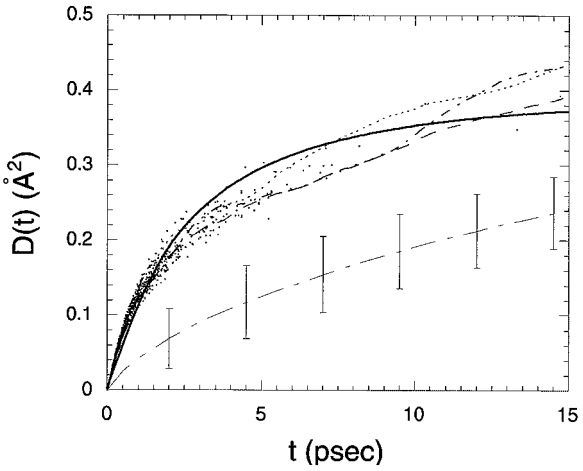


Fig. 4. Mean-square fluctuation (MSF) against duration time. $D^k(\tau_k)$ (dots), theoretical curve obtained by Langevin equation (thick solid line), $D^T(t)$ (three thin broken lines to the top), and $D^R(\tau)$ (thin dot-dashed line with error bars) are shown.

sensitive to differences of energy surface involved with low frequency modes. $D^k(\tau_k)$, which is the MSF during the residence time $\tau_k$ in the $k$th rotamer state, is identical with the trace of the matrix $\mathbf{R}^k$. $D^k(\tau_k)$ values determined for the 1,172 rotamer states are shown as a function of $\tau_k$ in Figure 4. These values are distributed in the very narrow range. This confirms the mutual similarity of the local energy landscape of these conformational substates.

Now we compare $D^k(\tau_k)$ with a curve determined theoretically. As shown before, autocorrelation functions determined by Langevin equation are in good agreement with autocorrelation functions directly calculated from MD. In the case of one-dimensional oscillator, MSF obtained by Langevin equation, $D^{lg}(\tau)$, is given by Equation B9 in Appendix B. Here total MSF is assumed to be given as a sum of $D^{lg}(\tau)$ over all the principal modes of the 683rd rotamer state. For each principal mode, angular frequency $\omega_0$ is set to $(k_BT/\lambda_m^{683})^{1/2}$ and a uniform value of 5 cm$^{-1}$ is assumed for the friction coefficient $\gamma$. As seen in Figure 4, the theoretical curve is in good agreement with $D^k(\tau_k)$. Therefore, it is concluded that Langevin equation with harmonic energy surface is a good model for atomic fluctuation in rotamer states.

The thick solid line in Figure 4 for the theoretical MSF curve indicates how $D(t)$ approaches its equilibrium value. Each dot in Figure 4 is placed at a time when a protein rotamer state undergoes a transition. Therefore, we see that the state point leaves a catchment region most of the time before equilibration of $D(t)$ is attained.

To study the effect of transitions between rotamer states, $D(\tau)$ curves obtained for three different periods, during which several rotamer transitions (T) take place, $D^T(t)$, are also shown. They are also in the range of intra-rotamer MSF. This will be interpreted

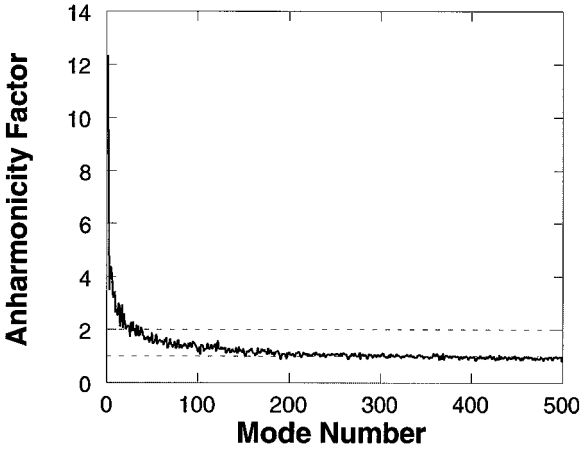


Fig. 5. Anharmonicity factor against the mode number of principal modes.

as follows. Most of the rotamer transitions involve dihedral-angle transitions in side chains. Axes of these transitions are expected to be nearly orthogonal to global motions which dominantly determine MSF. In such cases, rotamer transitions do not significantly change time-dependence of MSF.

In Figure 4, a function, $D^R(t)$, is also shown. $D^R(t)$ is a trace of the second-moment matrix $\mathbf{S}^R$ for time duration $t$. In other words, $D^R(t)$ represents a MSF originated only from rotamer-state transitions. In this time range, increasing rate of this curve is much smaller than that of intra-rotamer motions, $D^k(\tau_k)$. This confirms why difference between $D^T(t)$ and $D^k(\tau_k)$ was small. It should be also noted that error bar of the $D^R(t)$ is very large compared with the distributions of $D^k(\tau_k)$. This is due to stochastic nature of transition events.

### C. Classification of Principal Modes

Principal modes are classified by using two measures, anharmonicity factor of Equation 20 and deviation of $P_m^P(q_m^P)$ from gaussian distribution of Equation 23. The number of atoms in human lysozyme, $N$, is 2,041. The number of internal degrees of freedoms is $(3N - 6) = 6{,}117$.

In Figure 5, anharmonicity factor is shown. Anharmonicity factor is greater than two for the first 30 modes, greater than unity for the first 300 modes, and around unity for higher numbered modes. This means that anharmonic modes (1st $\sim$ 300th modes) and harmonic modes (301st $\sim$ 6,117th modes) are clearly distinguished by mode number. This has been already demonstrated in our previous paper in the case of bpti.[16] Therefore, the clear separation of anharmonic and harmonic modes by mode number appears to be a general property valid not only in bpti and lysozyme but also in any other globular proteins.

Examples of probability distribution functions are shown in Figure 6. Those of the first 30 modes have

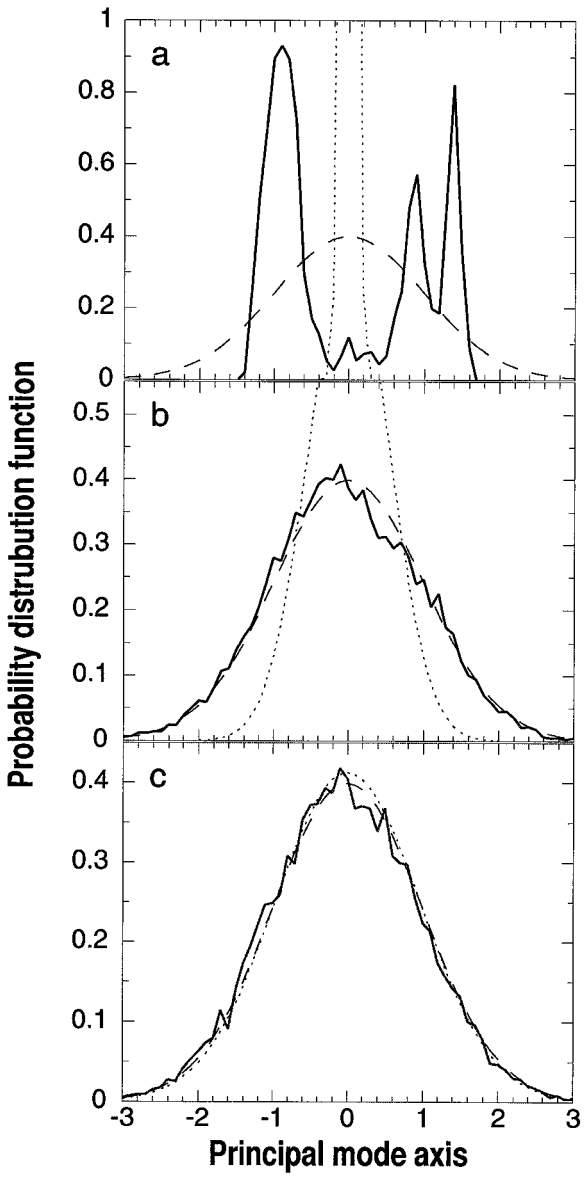


Fig. 6. Probability distribution functions along the **(a)** 1st, **(b)**, 50th, and **(c)** 500th principal modes. $P_m^{\alpha}(q_m^{\alpha})$ (solid lines), $P_m^{\alpha,gau}(q_m^{\alpha})$ (broken lines), and $P_m^{\alpha,har}(q_m^{\alpha})$ (dotted lines), respectively. Abscissa is scaled by root-mean-square fluctuation along each axis, (a) $(\zeta_1^P)^{1/2} = 0.76$ Å, (b) $(\zeta_{50}^P)^{1/2} = 0.057$ Å, and (c) $(\zeta_{500}^P)^{1/2} = 0.009$ Å, respectively.

multiple peaks as shown in Figure 6a. Especially, the peaks of the first four modes with root-mean-square fluctuation (RMSF) of 0.76, 0.50, 0.39, and 0.31 Å, respectively, are clearly separated. These four modes combined have 53.6% contribution to total MSF of the protein. Those of the 31st $\sim$ 300th modes have a gaussian-like single peak. However, as expected from the fact, $\rho_m^P > 1$, those are different from $P_m^{P,har}(q_m^P)$. Those of the 301st or higher-numbered modes are gaussians as expected from $P_m^{P,har}(q_m^P)$.

Considering anharmonicity factors and distribution functions, principal modes can be classified into

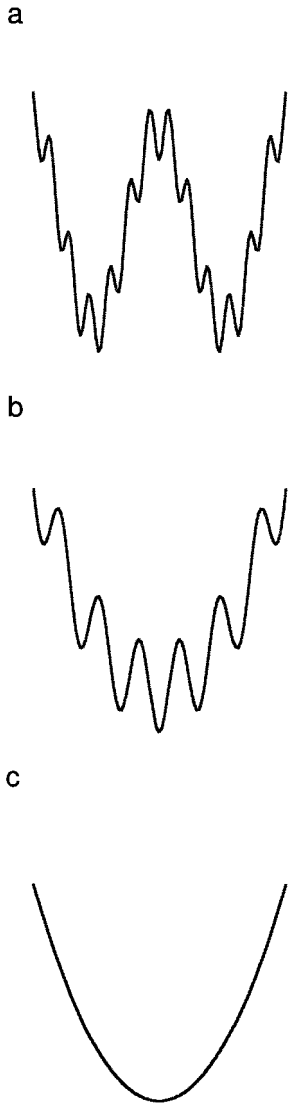


Fig. 7. Free energy surface is schematically shown for **(a)** multiply-hierarchical, **(b)** singly-hierarchical, and **(c)** harmonic modes, respectively.

three categories, i.e., multiply-hierarchical mode, singly-hierarchical mode, and harmonic mode. The higher numbered (301st ~ 6117th) modes are characterized by the value of unity of the anharmonicity factor, and by the gaussian distribution functions. This means that the energy surfaces for these modes are simply harmonic. Hence these modes are termed harmonic.

The 31st ~ 300th modes are gaussian-like in distribution functions, but with the anharmonicity factor greater than unity. This means that the energy surface has a harmonic envelope but with a multiple number of local minima. This situation is illustrated in Figure 7b. Along the axes of these modes, there should be a number of local minima arranged in a single coarse-grained higher-level minimum. In this sense, these modes are termed singly-hierarchical modes.

The 1st ~ 30th modes are termed multiply-hierarchical modes, because local minima are arranged in a multiple number of coarse-grained higher-level minima as illustrated in Figure 7a. The positions of the multiple peaks seen along multiply-hierarchical modes as in Figure 6a are understood as the position of the "higher level" conformational substates. This will be explained further in Section E. It should be noted that the levels of conformational states are ordered from the smallest energy level to higher in this paper and vice versa in the paper by Frauenfelder et al.[1]

Characteristics of these three types of modes are shown in Table II. Multiply-hierarchical modes dominantly contribute to total MSF. They undergo global collective motion. Multiply-hierarchical modes are related to dihedral angle fluctuation of both main and side chains. When the protein moves along these axes, intra-molecular packing topology[8] changes significantly. This means that these modes are involved with rearrangement of contacting atom pairs. The first principal mode is a typical hinge-bending motion. Singly-hierarchical modes are mainly related to dihedral angle fluctuation of side chains. Harmonic modes involve local motions that do not contribute significantly to total MSF.

### D. Relation Among Normal Modes, Principal Modes, JAM(I) Modes, and JAM(R) Modes

JAM(I) and JAM(R) mode analyses are carried out by the procedure described in Materials and Methods.

In JAM(I) mode analysis, matrix $\mathbf{S}^I$ should mean the contribution to the second-moment matrix from jumps among minima. Then, it is positively definite. However, it is guaranteed so only when state point stays in each minimum for a long enough time (see Equations A3 and A6). Because this may not be satisfied in the real system, $\mathbf{S}^I$ may have a number of negative eigenvalues. If protein rotamer states are considered as conformational substates, the equilibration conditions of Equations A3 and A6 are not well satisfied as shown in Figure 4. In our calculation, it turned out that most of the modes have positive or nearly-vanishing eigenvalues. Only twenty modes have non-negligible negative eigenvalues, but their magnitudes are relatively small. Judging from this, the matrix $\mathbf{S}^I$ satisfies the positive-definite condition not perfectly but fairly well. This means that, although the conditions of Equations A3 and A6 are not well satisfied, JAM(I) analysis works fairly well.

Now, we examine the relations among four coordinate systems, normal mode, principal mode, JAM(I) mode, and JAM(R) mode. The similarity of modes can be measured by coefficient $c_{mn}^{\alpha\beta}$ defined by Equation 16. Their values are shown in Figure 8. As clearly seen in Figure 8a and b, multiply-hierarchical and singly-hierarchical principal modes are closely

**TABLE II. Principal Mode Category**

| | Multiply-hierarchical mode | Singly-hierarchical mode | Harmonic mode |
|---|---|---|---|
| Principal mode number | 1–30 | 31–300 | 301–6117 |
| Probability distribution | Non-gaussian, multiple-peak | Gaussian-like, single-peak | Gaussian |
| Anharmonicity factor | $> 2$ | $> 1$ | $\sim 1$ |
| Percent in number of modes | 0.5 | 4.5 | 95.0 |
| Percent in total MSF | 82.1 | 15.0 | 2.9 |
| RMSF (Å)[a] | $> 0.08$ | $0.08 \sim 0.015$ | $0.015 >$ |
| Packing topology | Significant change | No change | No change |
| Fluctuation of dihedral angles | Large change in both main chains and side chains | Large change mainly in surface side chains | Small change |

[a]Root-mean-square fluctuation.

related to JAM(I) and JAM(R) modes of corresponding mode numbers. Especially multiply-hierarchical principal modes, whose $c^{PI}_{mn}$ and $c^{PR}_{mn}$ are larger than 0.5, can be understood as having one-to-one correspondence to JAM modes of the equal mode numbers. This is the reason why multiply- and singly-hierarchical modes are distinguished from harmonic modes in principal component analysis as shown in Section C. Multiply- and singly-hierarchical modes, put together, were termed anharmonic modes in the previous paper.[16] The fact that amplitude of anharmonic motions is much greater than that of harmonic modes underlies the effectiveness of principal component analysis in extracting anharmonic motions. Correspondence between principal and JAM(R) modes is slightly better than that between principal and JAM(I) modes. Although we do not show a figure, JAM(I) and JAM(R) also have clear one-to-one correspondence in the multiply-hierarchical region. The first 30 JAM(I) and JAM(R) modes are also considered as the multiply-hierarchical modes because of the clear one-to-one correspondence.

To the contrary, normal modes do not have clear one-to-one correspondence to principal modes and to JAM modes although there are some correlations as shown in Figure 8c. This explains the fact that projection to normal mode space is not an effective way of analyzing anharmonic motions as we have shown previously.[16] To further understand the relation between normal mode and JAM mode, subspace spanned by normal modes will be compared with JAM mode in Section F of Results.

### E. Hierarchy of Conformational Substates

In the first part of this section, we will show the characteristics of singly-hierarchical modes, and multiply-hierarchical modes. In order to understand the relation between these two kinds of modes, we divide the totally 1,000 psec trajectory into 10 parts, each of which consists of a 100 psec trajectory. First, average coordinates for each 100 psec trajectory are determined. In Figure 9a, these coordinates are projected onto the two-dimensional subspace spanned by the 1st and 100th principal modes. As shown in this figure, singly-hierarchical modes generally have no correlation with multiply-hierarchical modes. Although we do not show here, there are some correlations between some pairs of multiply-hierarchical modes. Probability distributions, $P^P_{100}(q^P_{100})$, along the axis of the 100th principal mode of the trajectories of each 100 psec duration are given by a gaussian distribution which is almost identical with that for the total length of the 1000-psec trajectory shown in Figure 9c. This means that distribution along the singly-hierarchical mode converges to its equilibrium form within 100 psec, independent of the motion along the multiply-hierarchical mode.

In Figure 9c, a distribution function of the protein rotamer states along the 100th principal mode is also shown. Because the position of each protein rotamer state should correspond to a local minimum, this distribution function reflects the shape of the envelope of complex energy surface. This function is also given by a gaussian distribution with a slightly smaller standard deviation compared with that of $P^P_{100}(q^P_{100})$. This means that the envelope is also harmonic. As was already shown, protein rotamer states are mutually similar and almost harmonic. Therefore, the situation corresponds to the case schematically illustrated in Figure 7b, i.e., protein rotamer states, 1st level conformational substates, are distributed on a harmonic envelope, which is understood as the 2nd level conformational substate.

In Figure 9b, $P^P_1(q^P_1)$ and the distribution function of the protein rotamer states along the 1st principal mode is shown. Judging from these distributions and from projections of average coordinate, it is concluded that relatively high energy barriers exist along this axis. In addition to the fast (1st level) transitions (taking place in a $\sim$1 psec time range) between protein rotamer states, two additional levels of transitions are observed along this axis. One is small-amplitude (2nd level) transitions taking place in the time range of $\sim$100 psec, and the other is large-amplitude transitions taking place in the time range longer than 400 psec. Along this type of axes,

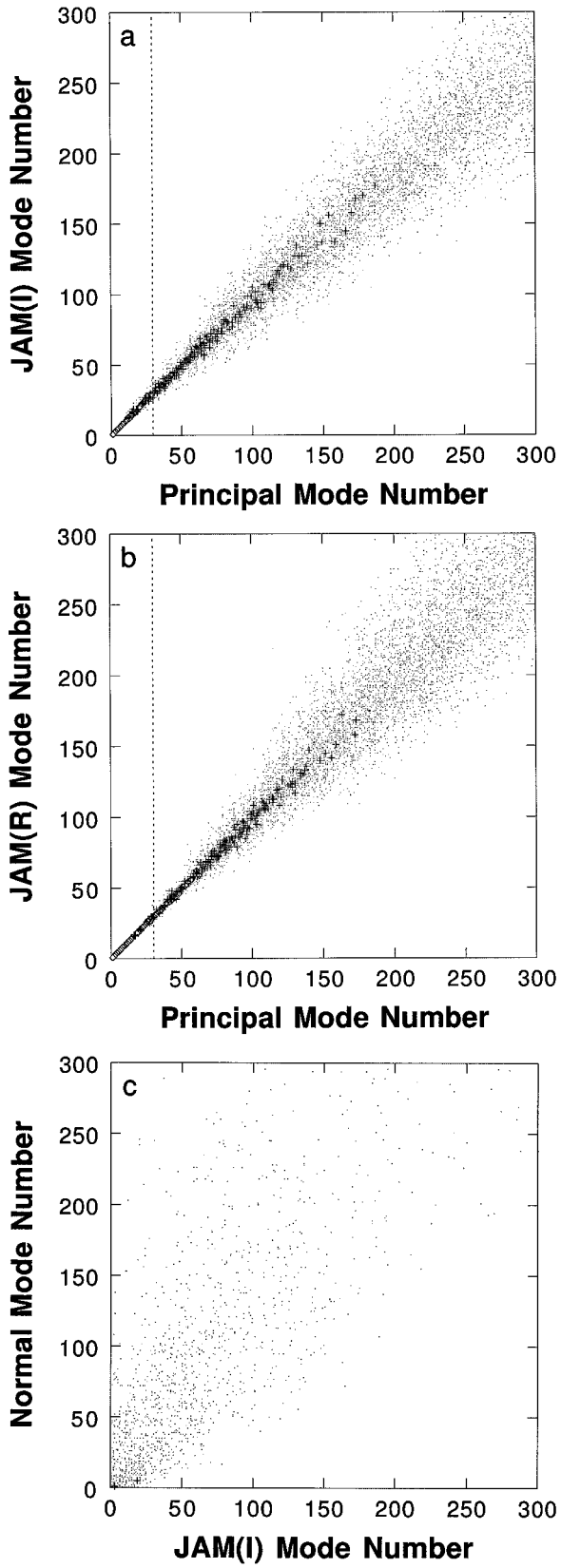


Fig. 8. Coefficients $c_{mn}^{\alpha\beta}$ between **(a)** principal and JAM(I) modes, **(b)** principal and JAM(R) modes, and **(c)** JAM(I) and normal modes are shown. The symbols, $\diamond$, $+$, and $\cdot$, represent that $c_{mn}^{\alpha\beta}$ is in the range, $1.0 \geq c_{mn}^{\alpha\beta} > 0.5$, $0.5 \geq c_{mn}^{\alpha\beta} > 0.1$, and $0.1 \geq c_{mn}^{\alpha\beta} > 0.01$, respectively. Thin vertical dotted line is plotted at the 30th principal mode, below which modes are multiply-hierarchical.

conformational substates are multiply hierarchical as illustrated in Figure 7a.

We now discuss the effects of hierarchical nature of the energy surface on residence time and MSF. Residence time in the 1st level catchment regions was given as residence time in protein rotamer states shown in Figure 1. As described in Section C, peaks of the distribution function along the first four modes are clearly separated. Therefore in these four modes we clearly observe the higher-level catchment regions. For these first four modes, principal modes, JAM(I) modes, and JAM(R) modes of corresponding mode number are identical, i.e., $c_{mm}^{\alpha\beta} \cong 1$. Residence times in the higher level catchment regions are shown in Table III. Residence times in the higher level catchment regions are in the time scale of $\sim$100 psec. We can observe an effect of such slow transitions also in time evolution of MSF. In Figure 10, $D^R(t)$ is again shown. Characteristics of $D^R(t)$ curve in the time range of $\geq$200 psec, is different from that in the range of $0 \sim 200$ psec. The longer time range behavior reflects JAM motions among the higher-level catchment regions existing in the subspace spanned by the first four multiply-hierarchical modes.

### F. Time Evolution of JAM Space

In this section, we investigate the time dependence of subspace spanned by JAM(R) modes. To do this, ten JAM(R) analyses are carried out for trajectories of the first 100, 200, . . . , 900, and 1000 psec of the totally 1000 psec MD. We designate the analysis done for the first t psec trajectory as JAM(R(t)) analysis. Firstly, let us consider the case where lower-numbered JAM(R(t)) eigenvectors are expressed as linear combinations of low-frequency normal-mode eigenvectors. As already shown in Figure 8c, normal modes do not have clear one-to-one correspondence with JAM mode. $P_m^{R(t)N}$ values given by the linear combination of the lowest 9 ($<10$ cm$^{-1}$), 203 ($<50$ cm$^{-1}$), 473 ($<100$ cm$^{-1}$), 672 ($<150$ cm$^{-1}$) normal modes are examined. Those determined for the first 100-psec trajectory (shown by superscript R(100)) are shown in Figure 11. Appearance of the plot is found essentially independent of the time duration of MD. $P_m^{R(100)N}$ values obtained by using the 9 lowest normal modes are relatively small. In the range of mode numbers shown in the figure, $P_m^{R(100)N}$ values obtained by using the 672 lowest-frequency normal modes are greater than 0.8. This means that the first 100 JAM(R) modes are well expressed as linear combinations of normal modes with frequen-

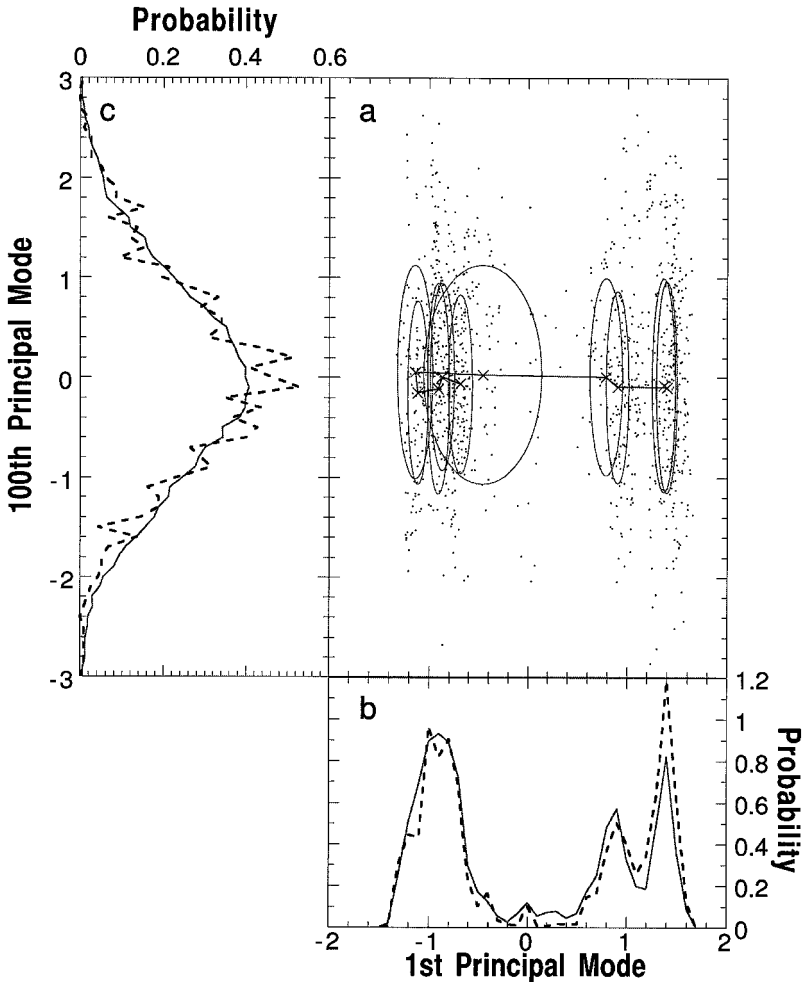


Fig. 9. **(a)** Projection of the 1172 protein rotamer states onto two-dimensional subspace spanned by the 1st and 100th principal modes shown by dots. The symbols '×' represent projections of coordinates of conformation averaged over every 100 psec trajectory. Ellipses represent root-mean-square fluctuations (RMSF) during each 100 psec trajectory along the 1st and 100th principal modes. **(b)** Probability distribution functions $P_m^P(q_m^P)$ (solid lines) and those for protein rotamer states (broken lines) for the 1st principal mode and **(c)** for the 100th principal mode, respectively. Abscissa of (b) and ordinate of (c) are scaled by RMSF, (b) $(\zeta_1^P)^{1/2} = 0.76$ Å and (c) $(\zeta_{100}^P)^{1/2} = 0.035$ Å, respectively.

**TABLE III. Residence Time in Conformational Substates**

| Principal mode number | Residence time in the catchment region of the higher level (psec) |
|---|---|
| 1 | >400 |
| 2 | ~250 |
| 3 | ~250 |
| 4 | ~200 |

cies lower than 150 $cm^{-1}$. It should be noted that normal modes, whose angular frequencies are smaller than 150 $cm^{-1}$, undergo non-local collective motions.[46,47] Therefore, the subspace in which inter-substate jumps take place is almost identical with the subspace spanned by the non-local collective normal modes.

In the same way, JAM(R(t)) modes determined by the first t psec MD trajectory are expressed by linear combination of JAM(R(200)) mode determined by the first 200 psec MD trajectory. Examples of $P_m^{R(t)R(200)}$ values are shown in Figure 12. Subspace

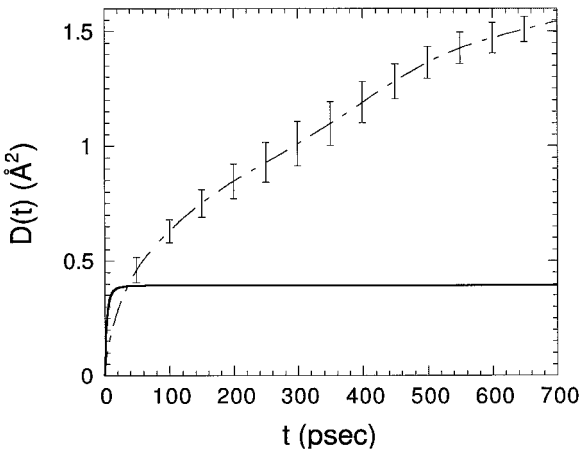


Fig. 10. Equivalence of Figure 4 in a different time scale. In this figure, MSF of JAM(R), $D^R(t)$ (thin dot-dashed line with error bars) and theoretical MSF obtained by Langevin equation (thick solid line) are shown.

spanned by the first 30 JAM(R(200)) modes are not well conserved during 1 nsec MD as shown by examples in Figure 12a. Therefore, the subspace

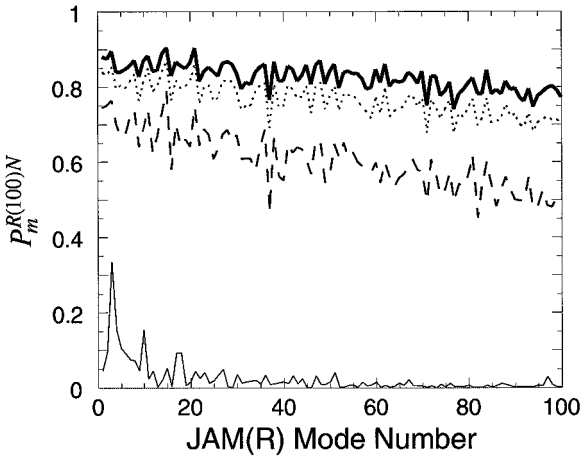


Fig. 11. Coefficients $P_m^{R(100)N}$ against the mode number of JAM(R(100)), when JAM (R(100)) mode eigenvectors are expressed as linear combination of low-frequency normal modes. $P_m^{R(100)N}$ determined by linear combinations of the first 9 (thin solid line), 203 (broken line), 473 (dotted line), and 672 normal modes (thick solid line) are shown. Frequency ranges of these normal modes are less than 10, 50, 100, and 150 $cm^{-1}$, respectively.

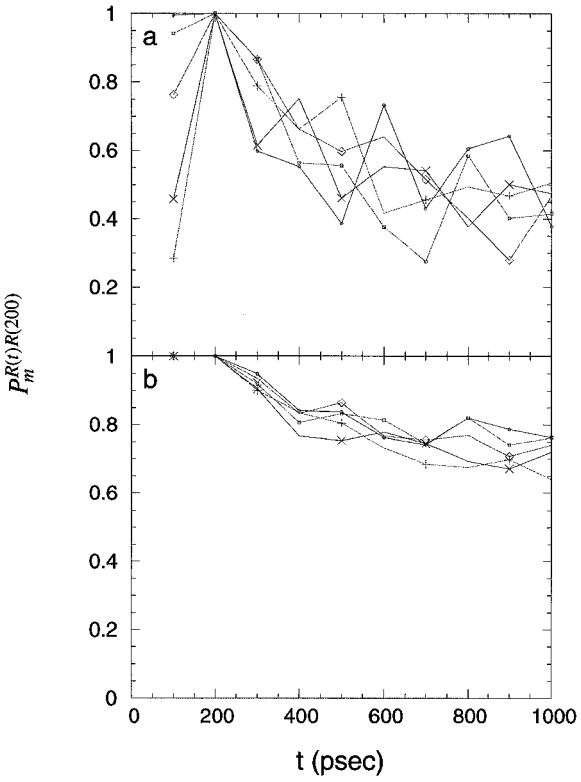


Fig. 12. Coefficients $P_m^{R(t)R(200)}$ against time, when JAM(R(t)) mode eigenvectors are expressed as linear combination of JAM(R(200)) modes determined by the first 200 psec trajectory. Examples of $P_m^{R(t)R(200)}$ determined by linear combinations of the first **(a)** 30 and **(b)** 300 JAM(R(200)) modes are shown. $P_m^{R(t)R(200)}$, whose m = (a) 5 (○), 10 (□), 15 (◇), 20 (×), 25 (+), (b) 10 (○), 30 (□), 50 (◇), 70 (×), 90 (+), are shown, respectively.

spanned by the first 30 JAM(R(200)) modes are rather different from the 30 multiply-hierarchical modes (see Table II) determined by 1 nsec MD. This means that multiply-hierarchical modes cannot be well defined by shorter MD. When the first 300 JAM(R(t)) modes determined by 1 nsec MD are expressed by the linear combinations of the first 300 JAM(R(200), $P_m^{R(t)R(200)}$ values are greater than ~0.7, independent on time t. Examples of values are shown in Figure 12b. Therefore, the subspace spanned by the 300 hierarchical modes (see Table II) are well conserved within the time range from 200 to 1000 psec. In other words, this subspace, which corresponds to what is called "important subspace"[48,49] appears to be time independent. Balsera et al. have claimed that principal component analysis does not reveal any reliable information on time scales that are not actually sampled.[71] However, the spirit of principal component analysis and JAM model exists in determining "important subspace," not in predicting slower motions which are not observed within the time range of simulation.

We can summarize the results of this section as follows. The subspace in which JAM motions take place can be identified by short (e.g., 200 psec) MD simulation. However, hierarchical conformational energy surface can be explored only by longer duration MD simulation.

## DISCUSSION

### Solvent Effect on Protein Motion

In this paper, we have shown that power spectral density in each rotamer state (conformational substates of the 1st level) is very similar to the number density of normal modes in vacuum (Figure 3). This means that solvent effect on the energy surface of the 1st-level conformational substates is relatively small. We think that this conclusion would be valid, independent of proteins.

However, friction effect on low-frequency collective motions is evidently protein dependent. Friction coefficients obtained by the method described in Reference 14 are 20 ± 10, 47 ± 10, and 100 ± 10 $cm^{-1}$ for human lysozyme (unpublished result), bpti,[14] and melittin,[12] whose molecular weights are 14,700, 6,520, and 2,580, respectively. This tendency against size of proteins is qualitatively explained by the relation $\gamma = 6\pi a\eta/m$, which is known as Stokes-Einstein law. Solvent viscosity $\eta$ is constant, and the factor $a/m \propto m^{-2/3}$ decreases rapidly as protein size increases.

In addition, it should be also noted that cut-off approximation employed in the MD simulation affects the magnitude of friction coefficient. In the present paper where no cut-off approximation is made, friction coefficient is around 5 $cm^{-1}$, whereas it has been 20 $cm^{-1}$ when cut-off approximation has been made in preceding unpublished work. In the

present case of low friction, frequency distribution determined from MD trajectory does not differ appreciably from that of normal mode as shown in Figure 3. However, frequency distribution in low frequency range should differ significantly from normal mode in cases of relatively high friction, as expected from Equation B7. It has been shown in fact in References 12 and 14. Various cut-off approximations, which produce large unphysical force at the cut-off distance, are known to affect the frequency distribution of normal mode in low frequency range.[50] Therefore, it is essential to employ highly accurate software package for MD.

In the principal component and JAM analyses, we did not pay attention to solvent degrees of freedom explicitly. In the case of polypeptide melittin, conformational transitions along the axes of the two largest-amplitude principal modes have been found to involve changes of hydration structure.[51] It has been shown also that some energy barriers on free energy surface are originated from hydration.[52] Although we did not analyze hydration structural change directly in this paper, we have indirect evidence that JAM motions have some correlations with hydration structural change. When a large-amplitude transition along the 1st JAM mode occurred, concurrent global translational and rotational motions were also observed. This means that water molecules around the lysozyme molecule move together with the lysozyme. Such motions are expected to involve changes of hydration structure as well. Therefore, we think that large JAM motions observed in this paper have some correlations with solvent motion.

### Anharmonicity of Protein Dynamics and Hierarchy of Conformational Substates

Protein dynamics has both harmonic and anharmonic aspects.[5,53] This has been explained by the fact that most of the degrees of freedom undergo with harmonic motions, whereas a small number of degrees of freedom undergo anharmonic motions which have dominant contribution to protein fluctuation.[15–17] In this paper, we clearly showed that the latter anharmonic motions are governed by hierarchical structure of energy surface along the axes of these motions.

Conformational substates are also characterized by their barrier heights between them. We can roughly estimate barrier heights among minima as follows. This can be done by using effective free energy $\mu_m^\alpha(q_m^\alpha)$ in Equation 22. Since motion along each axis is not completely independent, estimated barrier heights are not highly reliable. However, to give a rough order of barrier heights, it is still worth while estimating effective barrier heights this way. From a function obtained by smoothing of $P_m^I(q_m^I)$, we estimated barrier heights among minima. Result obtained for the multiply-hierarchical modes, the first 30 JAM(I) modes, is shown in Figure 13. Two

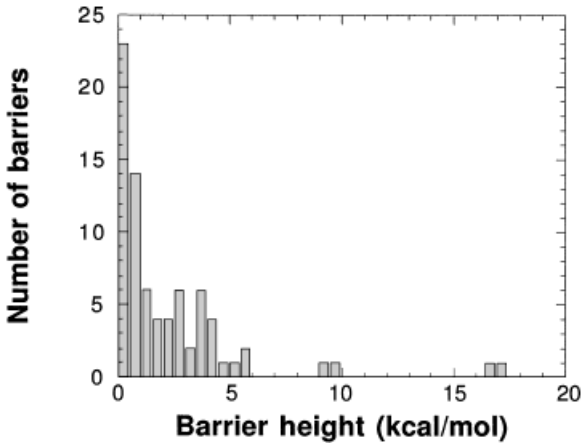


Fig. 13. Histogram of barrier heights in the free energy profile of the first 30 JAM(I) modes.

peaks around 17 kcal/mol in Figure 13 represent barrier heights of the higher level on the 1st JAM(I) mode. Two peaks around 9.5 kcal/mol represent barrier heights of the higher level on the 2nd JAM(I) mode.

Conformational substates are well studied for myoglobin experimentally. Conformational substates with various barrier heights have been reported.[1–3,54–56] Energy barriers found by hole burning experiments done at temperatures lower than 70 K,[54–56] which correspond to CS2 in Reference 1, are probably lower than those of the 1st level in the present paper. Note that we did not include both main-chain and side-chain terminal dihedral angles in the definitions of protein rotamer states. CS0, the highest level conformational substates found in myoglobin, is higher than those of the higher level in our treatment. Barrier heights of the 1st and higher levels shown in Figure 13 are comparable to activation enthalpy spectra of CS1, whose energy range is roughly from 1 to 9 kcal/mol.[1]

Straub and coworkers have also determined the distribution of the barrier heights from molecular dynamics of S-peptide of RNase A, bpti, and pentapeptide in vacuum.[57,58] Since their simulations were very short and done in vacuum, the results are quantitatively less reliable. However, they obtained barrier height distributions, which are similar to ours. Judging from this qualitative agreement of barrier-height distributions in various proteins, it is expected that the distribution curves of other proteins will be similar to that of human lysozyme. Unfortunately, their method cannot give concrete picture of conformations involved. An advantage of our method is that not only distribution of barrier heights but also specific jumping motions in the atomic detail can be described.

Leeson and Wiersma have pointed out the self-similarity of protein energy landscape[3] by also refer-

ring to temperature cycling hole burning.[54] They define self-similarity as follows. If a tier characterized by an average barrier height $E$ exists, higher or lower tiers with barrier heights of $\alpha E$, $\alpha^2 E$, . . . or $(1/\alpha)E$, $(1/\alpha)^2 E$, . . . also exist. Although we do not have a clear evidence of self-similarity in the present work, we think that the ideas of self-similarity and picture of hierarchical modes share a similar concept.

### Effective Use of JAM Analysis

In this section, we discuss the effectiveness of JAM analysis to investigate the protein energy landscapes, comparing with other methods. Recently theoretical approaches to study the protein energy landscape have been performed extensively.[4–17,27,51,52,57–62] In this kind of studies, conformational space is probed by using MD or Monte Carlo simulation. Then, the energy landscapes are analyzed mainly by clustering, energy minimization, projection of trajectory onto collective coordinate space, or combination of these methods.

Energy minimization is carried out by starting from many conformations sampled in the simulation to find the positions of local minima. Although this method has been found useful, it is limited to the cases in which simulations have been done in vacuum[4–9] or a small number of water molecules are included in the system.[10] We envisage two serious problems in energy minimization of a protein system with a large number of water molecules. (1) Since energy minimization of such system accompanies quenching of bulk water molecules as well as protein, many conformational substates originated also from variation of hydration structures will be observed. This will make the analysis intractably complicated. (2) Energy minimizations of protein with thousands of water molecules starting from many conformations needs a long computation time.

Projection of trajectory onto a collective coordinate space is a method more general and applicable also for proteins in water. As pointed out in Introduction, the key of this method is a choice of collective coordinate. The projection onto normal mode space is suitable for the analysis of intra-substate motions.[27] However, we have found that normal mode coordinate is less appropriate for the analysis of conformational substates.[16] Projection onto the principal mode space, which is determined by principal component analysis, is applicable to both intra- and inter-substate motions. Principal component analysis is widely used for the study of protein dynamics[12–16,63] and structure comparison.[64] It is essentially the same as the quasi-harmonic method[11,65] and essential dynamics.[17]

We have shown in this paper that JAM model has additional advantages. The most important advantage of JAM model is that protein motion is divided into intra-substate and inter-substate motions. To do this, we introduced strong physical assumptions in the actual applications of JAM model, JAM(I) and JAM(R) analyses. Considering the results of JAM analysis, we can conclude that assumptions we made are appropriate for proteins. By using JAM model, we could give answers to the four questions mentioned in Introduction. It is concluded that JAM model succeeded in exploring protein energy landscape.

We would suggest the proper use of principal component analysis and JAM analysis. Since both analyses give similar collective coordinate sets, we can choose one of these methods according to different purposes. If detailed analyses to give a physical picture of protein dynamics are necessary, JAM analysis is more suitable. If not, principal component analysis would be recommended because it is simpler and easier than JAM analysis.

### Relationship to Preceding Theoretical Works

Here we compare our present results with preceding theoretical works in which protein energy landscape has been investigated.[4–10,15–17,57,58,72,73] These simulations have been done in vacuum,[4–9,15,16,57,58,72] in partially solvated model,[10] and in water with 8 Å cut-off,[17] except for the recent work done in a crystal environment with Particle-Mesh Ewald method.[73] It has been reported that poor approximations in energy calculation cause large conformational change.[66–69] Therefore, artifacts originated from these approximations may be included in these cases. In this paper, we employed a software package in which one of the best sets of MD algorithms now available are incorporated. As already mentioned in Materials and Methods, not only highly accurate nonbonded calculation (CMM[32]) but also good boundary condition (SSBP[31]), and rigorous ensemble (Nose-Hoover[33,34]) are employed. High quality MD achieved by the combination of simulation algorithms employed here has been reported elsewhere.[30] We think quality of our simulation is much more improved than those in previous works.

Elber and Karplus have reported multiple conformational substates of myoglobin.[4] By using the scaled difference matrix, they have found that changes occur in one or few loop regions below 100 psec whereas changes occur globally in the slower time range. The latter seems to correspond to the transitions among the higher-level conformational substates found in this paper in human lysozyme along the first four multiply-hierarchical modes. By using the concept of ultrametricity and considering each on the minimized structures as replica, they have found three distinctive ranges of rms differences. The range (rms $>$ 1.5 Å) in which all structures are disjoint may correspond to the higher-level jumps along the first multiply-hierarchical mode, because a distance between a conformation represent-

ing a peak at 1.2 of Figure 6a and a conformation representing a peak at −1.0 is about 2.2 × $(\zeta_1^P)$ = 1.67 Å. Similarly, it appears that the range (1 < rms < 1.5 Å), in which there are sets of disjoint clusters, seems to correspond to the fluctuations along the 2nd, 3rd, and 4th multiply-hierarchical modes.

Noguti and Go have investigated hierarchical multiple substates of bpti.[5–9] Two types of transitions, a collective switch involving a few residues near surface and a collective switch occurring in the core of the bpti, have been found. Judging from the features shown in Table II, the latter directly corresponds to multiply hierarchical modes. The former is similar to singly hierarchical modes. However, the former involves the packing topology change whereas singly-hierarchical modes do not involve this change. Probably, a collective switch near surface corresponds to the modes, which are on the border of the multiply- and singly hierarchical modes.

In the analysis of 900 psec MD simulation of hen egg lysozyme done by Amadei et al.,[17] they have shown distribution similar to Figure 6 by using essential dynamics, which is equivalent to principal component analysis. What they call "essential subspace," which is a small subset of collective modes, is directly comparable to multiply hierarchical modes. Since essential dynamics is more oriented to efficient sampling, they have not gone into the nature of protein dynamics in detail. As shown in Section F of Results, subspace spanned by the multiply-hierarchical modes, which corresponds to essential subspace, is time dependent, whereas subspace spanned by multiply hierarchical modes plus singly hierarchical modes is time independent. Therefore, it would be better to use the latter subspace in sampling rather than just the "essential subspace."

Troyer and Cohen have studied energy landscape by energy minimization and clustering.[10] They have carried out clustering of conformations in various time scales, 1, 10, 100, and 1000 psec. As they have shown in Figure 3 of their paper, clusters of conformations, whose motions are faster than 1 psec, are detectable only by minimization. Whereas slower motions have been detected by both minimized and raw MD trajectories. Short time scale clusters (<1 psec) are comparable in time scale to rotamer states. As the highest level of clusters, they have found seven clusters. These clusters are separated by 0.7–1.0 Å, which are comparable to motions of the first four multiply hierarchical modes in human lysozyme. Multiply hierarchical modes in human lysozyme mainly involve global motions whereas inter-cluster motions in bpti have been found localized to a loop region involving the trypsin binding site.

Straub and Thirumalei have pointed out the existence of relatively fast transitions (<10 psec).[57] They determined these transitions from time dependence of reciprocal of the force metric. Slower motions greater than 15 psec were also found. Unfortunately, their method does not give concrete atomic picture of the transitions. Therefore, further comparison is difficult.

In our previous two works, we have examined energy landscape of bpti by using principal component analysis.[15,16] Using normal mode as a reference, we have found that anharmonic modes and harmonic modes are clearly separated by mode numbers. The reason has been already explained in Section D of Results in this paper. In the reference,[16] anharmonic modes have been classified to four types. The first three types of modes and the type 4 modes correspond to multiply and singly hierarchical modes, respectively. The type 1 mode (only one mode has been identified) corresponds to the first four multiply-hierarchical modes in lysozyme. Difference between type 2 and 3 is not seen in the present study. In the case of bpti, 200 principal modes among totally 1,698 modes have an anharmonic feature. The fraction of anharmonic modes determined by united-atom model employed in bpti, 11.7%, can be rescaled to 7.4% in all-atom model. It is comparable to the fraction of (multiply hierarchical modes plus singly hierarchical modes), 5.0% in human lysozyme. Judging from this, anharmonic (hierarchical) modes are expected to be a few percent of total modes in other proteins, too.

García and Harman have carried out 625-psec MD of CRP:$(cAMP)_2$ in vacuo and the principal component analysis of 418 $C^\alpha$ atoms.[72] Five principal modes (they call them MODCs instead of principal modes) of 1,224 total modes have "multimodal" feature. Transitions of these modes occurred in the order of 100 psec. These modes are comparable to the first four multiply hierarchical modes. Other modes, which are called "unimodal," seem to correspond to singly hierarchical and harmonic modes. García et al. have also carried out 5.1-nsec MD of the small protein crambin in the crystal environment.[73] From log-log plots of mean-square deviation from initial coordinate (msd) against time (note that the definition of their msd is different from that of MSF shown in Fig. 10), they found time evolution differs between the time range from 5 to 800 psec and that from 2 nsec to later. These two different behaviors may be due to the differences in the shapes of the envelopes of energy surface between lower and higher levels in hierarchy.

### Generality of the Results

Finally, we discuss the generality of the present results. Judging from the experimental and simulation results and considering the analogy to spin

glass, mutual similarity of the conformational substates has been expected for various proteins. However, we think these are the first results in which mutual similarity of the conformational substates are quantitatively shown in the atomic detail.

In the case of native human lysozyme, we showed that inter-substate motions occur in a small-dimensional conformational subspace. Considering the similarity between principal mode and JAM mode, we can discuss the generality of this result by examining the results of principal component analysis, which is widely used in the analysis of protein dynamics. In the cases where principal component analysis and its variations are employed, it is generally seen that only a small number of modes have anharmonic features. Therefore, it is expected that inter-substate motions generally occur in a small-dimensional subspace also in other proteins.

In this paper, we discussed the hierarchy of the conformational substates. In the cases discussed in the previous section, relatively large motions, which correspond to the first four multiply-hierarchical modes, have been found also in bpti and myoglobin. The motions found in bpti (multi-conformers in trypsin binding site) and human lysozyme (hinge bending motion) are thought to be related to functions. This suggests functional importance of hierarchical modes.

## CONCLUSION

We investigated energy landscape of human lysozyme by using principal component analysis and jumping-among-minima (JAM) model. We performed highly accurate molecular dynamics simulation and generated 1 nsec trajectory. Our purpose is to answer the four questions; (1) Are "conformational substates" mutually similar? (2) How are conformational substates distributed in the multi-dimensional conformational space? (3) Are conformational substates hierarchical? (4) How does the subspace, which contains conformational substates, evolves as a function of duration of time? To answer these questions, we employ principal component and JAM analyses. We apply two types of JAM modes, JAM(I) and JAM(R). In the former model, harmonicity of local energy surface is explicitly assumed without assigning positions of energy minima. In the latter, conformational substates are assumed to be rotamer states. By using the JAM models, we divide protein motions into intra-substate and inter-substate motions.

By analyzing rotamer states, we found that local energy surfaces of conformational substates are nearly harmonic and mutually similar. This is the answer to Question 1.

By using anharmonicity factor and shape of the probability distribution function, principal modes are classified into three types: (1) multiply-hierarchical modes, (2) singly-hierarchical modes, and (3) harmonic modes. The numbers of these three types of modes are 0.5%, 4.5%, and 95.0% of the total number of modes, respectively. However, multiply-hierarchical modes dominantly determine total atomic fluctuation of lysozyme.

Relation among principal modes, JAM(I) modes, JAM(R) modes, and normal modes is discussed. Multiply- and singly-hierarchical principal modes are understood as having one-to-one correspondence with JAM(I) modes and JAM(R) modes of the equal mode numbers. It is also concluded that the subspace, in which inter-substate motions occur, is spanned by the multiply- and singly-hierarchical modes, whose fraction of the number of modes is only 5% of the total number of modes. This is the answer to Question 2.

To answer Question 3, hierarchy of the substates is examined. We found at least two levels of inter-substate motions, the 1st and higher levels. Inter-substate motions of the 1st level, which are characterized by rotamer state transitions, have a feature of a first-order reaction. Inter-substate motions of the higher level, which are evidently seen also in principal component analysis, occur in the time scale of 200 psec or slower.

Time evolution of subspace spanned by the multiply- and singly-hierarchical modes is investigated to answer Question 4. In the time range of 100–1,000 psec, this subspace is well expressed by linear combination of low-frequency normal modes whose angular frequencies are smaller than 150 $cm^{-1}$. Anharmonic subspace determined by the first 200 psec MD is well conserved up to 1000 psec. However, distinction of multiply-hierarchical or singly-hierarchical, cannot be done by short simulation. Multiply-hierarchical feature of the modes becomes evident only after inter-substate transitions of higher level is observed.

Finally, we mention other possible applications of JAM model. Since the assumptions employed in JAM model is quite simple, it is generally applicable to the study of the protein energy landscapes. Therefore, applications of JAM model to the analyses of folding and unfolding simulations are possible.

JAM model could also be used as a tool to analyze experimental data. Protein dynamics is studied by experiments such as NMR, laser spectroscopy, neutron scattering, etc. Since JAM modes taking place in a small-dimensional space requires a small number of parameters for its description, these parameters could be well determined experimentally.

To interpret data obtained by these experiments, it is essential to employ a simple and good model. To give a physical picture of protein dynamics, a model should be good in the sense that it contains the essential feature of proteins. At the same time, the

model should not be too complicated because models with too many parameters are prone to overfitting. The JAM model, given by Equation 6, is a simple and good model. It is based only on two essential features of protein, intra-substate fluctuations and inter-substate transitions. We believe that the JAM model is applicable to analysis of experimental data and that it will be possible in the near future.

## ACKNOWLEDGMENTS

We express our thanks to Professor Benoît Roux for providing us FORTRAN subroutines for the SSBP calculation. MD simulation and analysis were done mainly by CRAY J916 in our group. A part of computation was done also at the Computer Centers of Kyoto University, Center for Promotion of Computational Science and Engineering of Japan Atomic Energy Research Institute, and Computer Center of the Institute for Molecular Science.

## REFERENCES

1. Frauenfelder, H., Sligar, S.G., Wolynes, P.G. The energy landscapes and motions of proteins. Science 254:1598–1603, 1991.
2. Frauenfelder, H., Wolynes, P.G. Biomolecules: Where the Physics of complexity and simplicity meet. Physics Today February 47:58–64, 1994.
3. Leeson, D.T., Wiersma, D.A. Looking into the energy landscape of myoglobin: Evidence for self-similarity. Nat. Struct. Biol. 2:848–851, 1995.
4. Elber, R., Karplus, M. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. Science 235:318–321, 1987.
5. Noguti, T., Go, N. Structural basis of hierarchical multiple substates of a protein. I: Introduction. Proteins 5:97–103, 1989.
6. Noguti, T., Go, N. Structural basis of hierarchical multiple substates of a protein. II: Monte Carlo simulation of native thermal fluctuations and energy minimization. Proteins 5:104–112, 1989.
7. Noguti, T., Go, N. Structural basis of hierarchical multiple substates of a protein. III: Side chain and main chain local conformations. Proteins 5:113–124, 1989.
8. Noguti, T., Go, N. Structural basis of hierarchical multiple substates of a protein. IV: Rearrangements in atom packing and local deformations. Proteins 5:125–131, 1989.
9. Noguti, T., Go, N. Structural basis of hierarchical multiple substates of a protein. V: Nonlocal deformations. Proteins 5:132–138, 1989.
10. Troyer, J.M., Cohen, F.E. Protein conformational landscapes: Energy minimization and clustering of a long molecular dynamics trajectory. Proteins 23:97–110, 1995.
11. Levy, R.M., Srinivasan, A.R., Olson, W.K., McCammon, J.A. Quasi-harmonic method for studying very low frequency modes in proteins. Biopolymers 23:1099–1112, 1984.
12. Kitao, A., Hirata, F., Go, N. The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulation of melittin in water and in vacuum. Chem. Phys. 158:447–472, 1991.
13. García, A.E. Large-amplitude nonlinear motions in proteins. Phys. Rev. Lett. 68:2696–2699, 1992.
14. Hayward, S., Kitao, A., Hirata, F., Go, N. Effect of solvent on collective motions in globular proteins. J. Mol. Biol. 234:1207–1217, 1993.
15. Hayward, S., Kitao, A., Go, N. Harmonic and anharmonic effects in the dynamics of BPTI: A normal mode analysis and principal component analysis. Protein Sci. 3:936–943, 1994.
16. Hayward, S., Kitao, A., Go, N. Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. Proteins 23:177–186, 1995.
17. Amadei, A., Linssen, A.B.M., Berendsen, H.J.C. Essential dynamics of proteins. Proteins 17:412–425, 1993.
18. Kidera, A., Inaka, K., Matsushima, M., Go, N. Normal mode refinement: Crystallographic refinement of protein dynamic structure. II. Application to human lysozyme. J. Mol. Biol. 225:477–486, 1992.
19. Zhang, X., Woxniak, J.A., Matthews, B.W. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J. Mol. Biol. 250:527–552, 1995.
20. McCammon, J.A., Gelin, B.R., Karplus, M., Wolynes, P.G. Hinge bending mode in lysozyme. Nature 262:325–326, 1976.
21. Brooks, B., Karplus, M. Normal modes for specific motions of macromolecules: Application to the hinge bending mode of lysozyme. Proc. Natl. Acad. Sci. USA 82:4995–4999, 1985.
22. Post, C.B., Brooks, B.R., Karplus, M., Dobson, C.M., Artymiuk, P.J., Cheetham, J.C., Phillips, D.C. Molecular dynamics simulations of native and substrate-bound lysozyme. J. Mol. Biol. 190:455–479, 1986.
23. Ichiye, T., Karplus, M. Anisotropy and anharmonicity of the atomic fluctuations in proteins: Analysis of a molecular dynamics simulation. Proteins 2:236–259, 1987.
24. Brooks III, C.L., Karplus, M. Solvent effects on protein motion and protein effects on solvent motion. J. Mol. Biol. 208:159–181, 1989.
25. Smith, L.J., Mark, A.E., Dobson, C.M., van Gunsteren, W.F. Comparison of MD simulations and NMR experiments for hen lysozyme. Analysis of local fluctuations, cooperative motions, and global changes. Biochemistry 34:10918–10931, 1995.
26. Gibrat, J., Go, N. Normal mode analysis of human lysozyme: Study of the relative motion of the two domains and characterization of the harmonic motion. Proteins 8:258–279, 1990.
27. Horiuchi, T., Go, N. Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: Human lysozyme. Proteins 10:106–116, 1991.
28. Arnold, G.E., Ornstein, R.L. A molecular dynamics simulation of bacteriophage T4 lysozyme. Protein Eng. 5:703–714, 1992.
29. Hayward, S., Kitao, A., Berendsen, H.J.C. Model-free methods of analyzing domain motions in proteins from simulations: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. Proteins 27: 425–437, 1997.
30. Sugita, Y., Kitao, A. Improved Protein Free Energy Calculation by More Accurate Treatment of Nonbonded Energy: Application to Chymotrypsin Inhibitor 2, V57A. Proteins 30:388–400, 1998.
31. Beglov, D., Roux, B. Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations. J. Chem. Phys. 100:9050–9063, 1994.
32. Ding, H.-Q., Karasawa, N., Goddard III, W.A. Atomic level simulations on million particles: The cell multipole method for Coulomb and London nonbond interactions. J. Chem. Phys. 97:4309–4315, 1992.
33. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. Mol. Phys. 52:255–268, 1984.
34. Hoover, W.G. Canonical dynamics: Equilibrium phase-space distributions. Phys. Rev. A 31:1695–1697, 1985.
35. Weiner, S.J., Kollman, P.A., Nguyen, D.T., Case, D.A. An all atom force field for simulations of proteins and nucleic acids. J. Comp. Chem. 7:230–252, 1986.
36. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79:926–935, 1983.
37. Gear, C.W. Numerical Initial Value Problems in Ordinary

Differential Equations. Chap. 7, Englewood Cliffs, New Jersey: Prentice-Hall Inc. 1971.
38. Kitao, A., Go, N. Conformational dynamics of polypeptides and proteins in the dihedral angle space and in the cartesian coordinate space: Normal mode analysis of decaalanine. J. Comp. Chem. 12:359–368, 1991.
39. Eckart, C. Some Studies Concerning Rotating Axes and Polyatomic Molecules. Phys. Rev. 47:552–558, 1935.
40. Kitao, A., Hayward, S., Go, N. Comparison of normal mode analyses on a small globular protein in dihedral angle space and cartesian coordinate space. Biophys. Chem. 52:107–114, 1994.
41. Tomimoto, M., Kitao, A., Go, N. Normal mode analysis of a nucleic acid with flexible furanose rings in dihedral angle space. Electro. J. Theor. Chem. 1:122–134, 1996.
42. Takano, K., Ogasahara, K., Kaneda, H. et al. Contribution of hydrophobic residues to the stability of human lysozyme: Calorimetric studies and X-ray structural analysis of the five isoleucine to valine mutants. J. Mol. Biol. 254:62–76, 1995.
43. Artymuik, P.J., Blake, C.C.F. Refinement of human lysozyme at 1.5Å resolution analysis of non-bonded and hydrogen-bonded interactions. J. Mol. Biol. 152:737–762, 1981.
44. Uhlenbeck, G.E., Ornstein, L.S. On the theory of the Brownian motion. Phys. Rev. 36:823–841, 1930.
45. Chandrasekhar, S. Stochastic problems in physics and astronomy. Reviews of Modern Physics 15:1–89, 1943.
46. Go, N., Noguti, T., Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc. Natl. Acad. Sci. USA 80:3696–3700, 1983.
47. Nishikawa, T., Go, N. Normal modes of vibration in bovine pancreatic inhibitor and its mechanical properties. Proteins 2:308–329, 1987.
48. Kidera, A., Go, N. Normal mode refinement: Crystallographic refinement of protein dynamic structure. I. Theory and test by simulated diffraction data. J. Mol. Biol. 225:457–475, 1992.
49. Hayward, S., Go, N. Collective variable description of native protein dynamics. Annu. Rev. Phys. Chem. 46:223–250, 1995.
50. Smith, J.C. Protein dynamics: comparison of simulations with inelastic neutron scattering experiments. Q. Rev. Biophys. 24:227–291, 1991.
51. Kitao, A., Hirata, F., Go, N. Effects of solvent on the conformation and the collective motions of a protein. 2. Structure of hydration in melittin. J. Phys. Chem. 97: 10223–10230, 1993.
52. Kitao, A., Hirata, F., Go, N. Effects of solvent on the conformation and the collective motions of a protein. 3. Free energy analysis by the extended RISM Theory. J. Phys. Chem. 97:10231–10235, 1993.
53. Go, N., Noguti, T. Structural basis of hierarchical multiple substates of a protein. Chemica Scripta 29A:151–164, 1989.
54. Shibata, Y., Kurita, A., Kushida, T. Structural relaxations in H2-substituted myoglobin observed by temperature-cycling hole burning. J. Chem. Phys. 104:4396–4405, 1996.
55. Kurita, A., Ohmukai, R., Kushida, T. Structural dynamics of proteins investigated by hole-burning spectroscopy. Journal of Luminescence 53:255–258, 1992.
56. Kurita, A., Shibata, Y., Kushida, T. Two-level systems in myoglobin probed by non-lorenzian hole broadening in temperature-cycling experiment. Phys. Rev. Lett. 74:4349–4352, 1995.
57. Straub, J.E., Thirumalai, D. Exploring the energy landscape in proteins. Proc. Natl. Acad. Sci. USA 90:809–813, 1993.
58. Straub, J.E., Choi, J.-K. Extracting the energy barrier distribution of a disordered system from the instantaneous normal mode density of states: Applications to peptides and proteins. J. Phys. Chem. 98:10978–10987, 1994.
59. Brooks, B.R., Janezic, D., Karplus, M. Harmonic analysis of large systems. I. Methodology. J. Comp. Chem. 16:1522–1542, 1995.
60. Janezic, D., Brooks, B.R. Harmonic analysis of large systems. II. Comparison of different protein models. J. Comp. Chem. 16:1543–1553, 1995.
61. Janezic, D., Veneble, R.M., Brooks, B.R. Harmonic analysis of large systems. III. Comparison with molecular dynamics. J. Comp. Chem. 16:1554–1566, 1995.
62. Roitberg, A., Gerber, R.B., Elber, R., Ratner, M.A. Anharmonic wave functions of proteins: Quantum self-consistent field calculations of BPTI. Science 268:1319–1322, 1995.
63. Basu, G., Kitao, A., Hirata, F., Go, N. A collective motion description of the $3_{10}$-/α-helix transition: Implications for a natural reaction coordinate. J. Am. Chem. Soc. 116:6307–6315, 1994.
64. Takahashi, K., Go, N. Conformational classification of short backbone fragments in globular proteins and its use for coding backbone conformations. Biophys. Chem. 47:163–178, 1993.
65. Levy, R.M., Rojas, O.L., Friesner, R.A. Quasi-harmonic method for calculating vibrational spectra from classical simulations on multidimensional anharmonic potential surfaces. J. Phys. Chem. 88:4233–4238, 1984.
66. Levitt, M., Sharon, R. Accurate simulation of protein dynamics in solution. Proc. Natl. Acad. Sci. USA 85:7557–7561, 1988.
67. Loncharich, R.J., Brooks, B.R. The effects of truncating long-range forces on protein dynamics. Proteins 6:32–45, 1989.
68. Saito, M. Molecular dynamics simulations of proteins in solution: Artifacts caused by the cutoff approximation. J. Chem. Phys. 101:4055–4061, 1994.
69. York, D.M., Wlodawer, A., Pedersen, L.G., Darden, T.A. Atomic-level accuracy in simulations of large protein crystals. Proc. Natl. Acad. Sci. USA 91:8715–8718, 1994.
70. Morikami, K., Nakai, T., Kidera, M., Saito, M., Nakamura, H. PRESTO (Protein Engineering SimulaTOr): A vectorized molecular mechamics program for biopolymers. Comput. Chem. 16:243–248, 1992.
71. Balsea, M.A., Wriggers, W. Oono, Y., Schulten, K. Principal component analysis and long time protein dynamics. J. Phys. Chem. 100:2567–2472, 1996.
72. García, A.E., Harman, J.G. Simulation of CRP:$(cAMP)_2$ in noncrystalline environments show a subunit transition from the open to the closed conformation, Protein Sci. 5:62–71, 1996.
73. García, A.E., Blumenfeld, R., Hummer, G., Krumhansl, J.A. Multi-basin dynamics of a protein in a crystal environment. Physica D 107:225–239, 1997.

## APPENDIX A

### JAM Model in an Idealized Case

First we introduce a set of normal mode eigenvectors at each minimum. Let $w_{ij}^k$ be Cartesian component of the $j$th eigenvector at the $k$th minimum. For simplicity, the superscript N, which stands for normal mode in Materials and Methods, is omitted in Appendix. We express an instantaneous value of $x_i(t)$ in reference to coordinate $x_i^k$ of a minimum, in the catchment region of which the instantaneous conformational point exists. Thus,

$$x_i(t) = x_i^k + \sum_j w_{ij}^k \sigma_j^k(t), \tag{A1}$$

where $\sigma_j^k(t)$ is the projection of deviation of $x_i(t)$ from the minimum $x_i^k$ onto the normal mode axis given by $w_{ij}^k$. We now consider an average of $x_i(t)$ over a period

of simulation. It is given by,

$$\langle x_i \rangle = \sum_k f_k x_i^k + \sum_{k,j} f_k w_{ij}^k \langle \sigma_j^k \rangle_k, \tag{A2}$$

where $f_k$ is the fraction of time of the period, during which the instantaneous state point stayed at the $k$th catchment region, and $\langle \sigma_j^k \rangle_k$ is the average of $\sigma_j^k(t)$ over that period.

At this point we introduce our first assumption;

$$\langle \sigma_j^k \rangle_k = 0. \tag{A3}$$

This equation can be justified by assuming that the state point stays in each catchment region for a significant period so that its projection on each of local normal mode axes assumes both positive and negative values at least a few times to average out to vanish. When the assumption of Equation A3 is made, Equation A2 reduces to,

$$\langle x_i \rangle = \sum_k f_k x_i^k. \tag{A4}$$

Then, second moment matrix of Equation 2 is given, by using Equation A3, as,

$$a_{ij} = \sum_k f_k (x_i^k - \langle x_i \rangle)(x_j^k - \langle x_j \rangle) + \sum_k f_k \sum_{l,m} w_{il}^k w_{ij}^k \langle \sigma_l^k \sigma_m^k \rangle_k. \tag{A5}$$

In the last expression $\langle \sigma_l^k \sigma_m^k \rangle_k$ is the average of $\sigma_l^k(t)$ $\sigma_m^k(t)$ over the period, during which the state point stays in $k$th catchment region. This corresponds to the expression of Equation 6 written in terms of normal mode variables. We now introduce our second assumption;

$$\langle \sigma_l^k \sigma_m^k \rangle_k = \delta_{lm} \frac{k_B T}{\omega_{kl}^2}, \tag{A6}$$

where $\delta_{lm} = 1$ for $l = m$, and $\delta_{lm} = 0$ for $l \neq m$, $\omega_{kl}$ the angular frequency of $l$th normal mode at $k$th minimum. This is an exact expression for long-enough pure harmonic motions. Therefore, the assumption of Equation A6 is valid when the state point stays in each catchment region for significant period (jumping-among-minima model) and motion in each catchment region is harmonic. By introducing Equation A6 into Equation A5, we have

$$a_{ij} = \sum_k f_k (x_i^k - \langle x_i \rangle)(x_j^k - \langle x_j \rangle) + \sum_{k,l} f_k w_{il}^k w_{jl}^k \frac{k_B T}{\omega_{kl}^2}. \tag{A7}$$

Let us consider an idealized case, where in Equation A7 $w_{il}^k$ and $\omega_{kl}$ do not depend on $k$. This means that eigenvectors and eigenfrequencies at all minima are the same. In this idealized case, Equation A7 reduces to Equation 7.

## APPENDIX B
## Correlation Function, Mode Number Density, and Mean-Square Fluctuation in Solution of Langevin Equation

Langevin equation in a harmonic potential is given by,

$$d^2x/dt^2 + \gamma\, dx/dt + \omega_0^2 x = A(t), \tag{B1}$$

where $x$, $\gamma$, $A(t)$, and $\omega_0$ are a coordinate, a friction constant, random force, and angular frequency of the oscillator. The solution of this equation is given by,[44,45]

$$x(t) = x_0 e^{-\gamma t/2}\left(\cosh \omega_1 t/2 + \frac{\gamma}{\omega_1} \sinh \omega_1 t/2\right) + \frac{2u_0}{\omega_1} e^{-\gamma t/2} \sinh \omega_1 t/2 + \frac{2}{\omega_1} \times \int_0^t d\xi [A(\xi) e^{-\gamma(t-\xi)t/2} \sinh \omega_1 (t-\xi)/2] \tag{B2}$$

where $x_0$ and $u_0$ are initial coordinate and initial velocity, respectively, and $\omega_1$ is defined by

$$\omega_1^2 = \gamma^2 - 4\omega_0^2. \tag{B3}$$

The averages of $A(\xi)$ and $A(\xi)A(\eta)$ over a restricted ensemble with a given set of the initial conditions, $x_0$ and $u_0$, are assumed to be,

$$\overline{A(\xi)} = 0, \overline{A(\xi)A(\eta)} = 2\gamma k_B T \delta(\xi - \eta), \tag{B4}$$

where $\delta(\xi - \eta)$ is a Dirac's delta function. By averaging the initial conditions over an equilibrium ensemble, we should have,

$$\overline{u_0} = 0, \overline{u_0^2} = k_B T, \overline{x_0^2} = k_B T/\omega_0^2. \tag{B5}$$

By using Equations B4 and B5, we obtain autocorrelation function of a coordinate, $C(t)$, as

$$C(t) \equiv \frac{\overline{x(t)x(0)}}{\overline{x(0)x(0)}} = e^{-\gamma t/2}\left(\cosh \omega_1 t/2 + \frac{\gamma}{\omega_1} \sinh \omega_1 t/2\right), \tag{B6}$$

where $\overline{\cdots}$ is an ensemble average and $\overline{x(0)x(0)} = \overline{x_0^2} = k_BT/\omega_0^2$. Then the mode number density is defined as the Fourier transform of the velocity autocorrelation function $\overline{u(t)u(0)}$, and can be obtained also from position autocorrelation function $\overline{x(t)x(0)}$ by,

$$F(\omega) = (k_BT/2\pi)\int_{-\infty}^{\infty} dt\,\overline{u(t)u(0)}\cos\,\omega t$$
$$= (k_BT\omega^2/2\pi)\int_{-\infty}^{\infty} dt\,\overline{x(t)x(0)}\cos\,\omega t$$
$$= \frac{1}{\pi}\left[\frac{\gamma\omega^2}{(\omega_0^2-\omega^2)^2+\gamma^2\omega^2}\right]. \quad \text{(B7)}$$

We define mean-square-fluctuation (MSF) during time period of $\tau$, $D(\tau)$, as,

$$D(\tau) = \frac{1}{\tau}\int_0^\tau dt\,(x(t))^2 - \frac{1}{\tau^2}\left(\int_0^\tau dt_1 x(t_1)\right)\left(\int_0^\tau dt_2 x(t_2)\right). \quad \text{(B8)}$$

In the case of the behavior expected for one-dimensional Langevin equation, MSF, $D^{lg}(\tau)$, is obtained by substituting Equation B2 into Equation B8 and by considering Eqs B4 and B5 as,

$$D^{lg}(\tau) = \frac{k_BT}{\omega_0^2}\left[1-\frac{1}{\tau^2}\left\{\int_0^\tau dte^{-\gamma t/2}\left(\cosh\,\omega_1 t/2 + \frac{\gamma}{\omega_1}\sinh\,\omega_1 t/2\right)\right\}^2 - \frac{1}{\tau^2}\frac{4\omega_0^2}{\omega_1^2}\left\{\int_0^\tau dte^{-\gamma t/2}\sinh\,\omega_1 t/2\right\}^2\right]$$
$$= \frac{k_BT}{\omega_0^2}\left[1-\frac{\gamma^2}{\omega_0^2(\omega_0\tau)^2}\left\{1-e^{-\gamma t/2}\left(\cosh\,\omega_1 t/2 + \frac{(\gamma+\omega_1)^2}{2\gamma\omega_1}\sinh\,\omega_1 t/2\right)\right\}^2 - \frac{1}{(\omega_0\tau)^2}\left\{1-e^{-\gamma t/2}\left(\cosh\,\omega_1 t/2 + \frac{\gamma}{\omega_1}\sinh\,\omega_1 t/2\right)\right\}^2\right]. \quad \text{(B9)}$$

The former integral form in Equation B9 is practically more useful in numerical calculation than the latter form in Equation B9. In the normal mode limit ($\gamma \to 0$), MSF is given by,

$$D^{nm}(\tau) = \frac{kT}{\omega_0^2}\left(1-\frac{2}{(\omega_0\tau)^2}(1-\cos\,\omega_0\tau)\right). \quad \text{(B10)}$$