Dynamics of essential collective motions in proteins: Theory

Maria Stepanova

National Institute for Nanotechnology, National Research Council of Canada, Department of Electrical and Computer Engineering, University of Alberta, 11421 Saskatchewan Drive, Edmonton, Alberta, Canada T6G 2M9

(Received 20 November 2006; revised manuscript received 23 July 2007; published 26 November 2007)

A general theoretical background is introduced for characterization of conformational motions in protein molecules, and for building reduced coarse-grained models of proteins, based on the statistical analysis of their phase trajectories. Using the projection operator technique, a system of coupled generalized Langevin equations is derived for essential collective coordinates, which are generated by principal component analysis of molecular dynamic trajectories. The number of essential degrees of freedom is not limited in the theory. An explicit analytic relation is established between the generalized Langevin equation for essential collective coordinates and that for the all-atom phase trajectory projected onto the subspace of essential collective degrees of freedom. The theory introduced is applied to identify correlated dynamic domains in a macromolecule and to construct coarse-grained models representing the conformational motions in a protein through a few interacting domains embedded in a dissipative medium. A rigorous theoretical background is provided for identification of dynamic correlated domains in a macromolecule. Examples of domain identification in protein G are given and employed to interpret NMR experiments. Challenges and potential outcomes of the theory are discussed.

DOI: 10.1103/PhysRevE.76.051918 PACS number(s): 87.15.He, 05.10.Gg, 87.15.Aa, 02.50.Sk

I. INTRODUCTION

Biological macromolecules show a tremendous range of various useful functionalities. The well-defined architecture of polymers combined with soft texture, their precise recognition power, and outreach diversity make biopolymers unique natural building blocks for targeted drug delivery, sensing, diagnostics, actuating systems, and molecular electronics. However, many biopolymer-based nanotechnologies are still in the stage of early infancy. A part of the reason for this is an insufficient understanding of structure and functionalities of organic macromolecules. Proteins are complex soft-matter systems containing thousands of atoms and apt to change their spatial conformation both spontaneously and through interactions with environment. One largely unsolved challenge for theoretical description and modeling is the need to account for many internal degrees of freedom in large polymer molecules. Currently, theoretical understanding of the conformational behaviors of proteins lags significantly behind the practical needs.

This paper introduces a general theoretical background for characterization and comparison of conformational motions in protein molecules. Section II briefly outlines the existing statistical technique to extract essential collective degrees of freedom from atomic trajectories; Sec. III introduces a general derivation of dynamic equations of motion for essential collective modes in a protein molecule; in Sec. IV, the introduced formalism is applied to define coarse-grained models representing the conformational motions in a protein through a few dynamic domains, with examples given for protein G; Sec. V discusses challenges and potential further developments of the theory; and Sec. VI summarizes the conclusions.

II. ATOMIC TRAJECTORIES, ESSENTIAL COLLECTIVE COORDINATES, AND FLUCTUATIONS

Trajectories of individual atoms in a macromolecule are provided, with a reasonable precision, by molecular dynamics simulations. Thus, positions of individual atoms in a system composed of a protein embedded in a solvent, and containing the total of N atoms, can be represented by one vector in the $3N$-dimensional phase space, $\vec{X}=(X_1,X_2,\ldots,X_{3N})$, were X_i are the coordinates of individual atoms. The time evolution of this vector provides the trajectory in the phase space, $\vec{X}(t)=(X_1(t),X_2(t),\ldots,X_{3N}(t))$, which represents the change of the conformation in the protein. However, capabilities to characterize and compare the various conformations from the molecular dynamics trajectories are rather limited. A part of the problem is that the information provided by direct molecular dynamics simulations is highly redundant and must be appropriately filtered in such a way that essential dynamic characteristics emerge.

Statistical ranking of complex dynamic systems is available through well-established techniques of the principal component analysis (PCA) [1]. When PCA is applied to the phase trajectory of a protein molecule [2–7], the covariance matrix is constructed,

$$c_{ij}=\langle[X_i(t)-\langle X_i\rangle][X_j(t)-\langle X_j\rangle]\rangle_{\text{traj}}, \tag{1}$$

where the averaging is over the entire trajectory [8]. For this covariance matrix, the normalized eigenvectors $\vec{E}^k=\{E_1^k,E_2^k,\ldots,E_{3N}^k\}$ and the eigenvalues σ^k $(k=1,2,\ldots,3N)$ are defined by

$$\sum_j c_{ij}E_j^k=\sigma^k E_i^k. \tag{2}$$

The eigenvectors $\vec{E}^1,\vec{E}^2,\ldots,\vec{E}^{3N}$ represent a set of $3N$ orthogonal collective degrees of freedom. One can consider the eigenvectors as the intrinsic coordinate frame in the phase space and project on them the phase trajectory $\vec{X}(t)$ [2]:

$$[\vec{X}(t)\cdot\vec{E}^k]=\sum_{i=1}^{3N}E_i^kX_i(t)=x^k(t),\quad k=1,2,\ldots,3N. \tag{3}$$

The functions $x^k(t)$ defined by Eq. (3) can be viewed as the collective coordinates that represent the conformational behavior of the protein molecule together with the solvent around it. The inverse operation provides the initial trajectories of the atoms, $X_i(t)$:

$$\sum_{k=1}^{3N}x^k(t)E_i^k=X_i(t),\quad i=1,2,\ldots,3N. \tag{4}$$

The eigenvalues defined by Eq. (2), σ^k, represent the mean-square displacements. The conventional approach is to rank the collective degrees of freedom according to the magnitude of the associated eigenvalues and to consider a truncated coordinate frame $\vec{E}^1,\vec{E}^2,\ldots,\vec{E}^{k_{\max}}$, $k_{\max}<3N$, which includes only those collective coordinates that have the highest magnitude of the displacements [2–7,9,10]. This truncated coordinate frame is also known at the essential degrees of freedom [5,6]. The complementary set of collective coordinates, $\vec{E}^{k_{\max}+1},\vec{E}^{k_{\max}+2},\ldots,\vec{E}^{3N}$, are interpreted as small-amplitude fluctuations. The essential degrees of freedom and those associated with fluctuations can be viewed as two orthogonal subspaces of $3N$-dimensional phase space. Accordingly, Eq. (4) can be replaced by a system of two complementary equations

$$\sum_{k=1}^{k_{\max}}x^k(t)E_i^k=X_i^E(t),$$

$$\sum_{k=k_{\max}+1}^{3N}x^k(t)E_i^k=X_i^{1-E}(t), \tag{5}$$

where the $X_i^E(t)$ represents the essential component of atomic trajectories and $X_i^{1-E}(t)$ represents the fluctuations. At every moment of time, the essential component and the fluctuation component of the atomic coordinates can be considered as orthogonal vectors in the phase space, $\vec{X}^E$ and $\vec{X}^{1-E}$, respectively, such that

$$\vec{X}=\vec{X}^E+\vec{X}^{1-E}. \tag{6}$$

This representation of the vector $\vec{X}$ by two orthogonal components can be efficiently accomplished using the Mori projection operator formalism [11]. Thus, consider the operator P that converts the vector $\vec{X}$ into $\vec{X}^E$ and the complementary operator $1-P$ that converts $\vec{X}$ into $\vec{X}^{1-E}$:

$$P\vec{X}=\vec{X}^E,$$

$$(1-P)\vec{X}=\vec{X}^{1-E}. \tag{7}$$

One can easily check that the operators P and $1-P$ are defined by

$$P\vec{X}=\sum_{k=1}^{k_{\max}}(\vec{E}^k\cdot\vec{X})\vec{E}^k,$$

$$(1-P)\vec{X}=\sum_{k=k_{\max}+1}^{3N}(\vec{E}^k\cdot\vec{X})\vec{E}^k. \tag{8}$$

From Eq. (8) it follows that the operators P and $1-P$ applied to the vector $\vec{X}$ can be interpreted as the geometrical projections of the vector $\vec{X}$ onto the subspace of the essential degrees of freedom and on the subspace of the fluctuations, respectively. This can be illustrated by the following simple examples:

$$P\vec{X}^E=\vec{X}^E,$$

$$P\vec{X}^{1-E}=0,$$

$$(1-P)\vec{X}^E=0,$$

$$(1-P)\vec{X}^{1-E}=\vec{X}^{1-E}. \tag{9}$$

Accordingly, the functions $\vec{X}^E(t)$ and $\vec{X}^{1-E}(t)$ can be viewed as projections of the phase trajectory $\vec{X}(t)$ onto the subspace of essential degrees of freedom and onto the subspace of fluctuations, respectively.

The truncated set of essential degrees of freedom has been extensively employed to study proteins [12], and such studies have provided valuable information about the geometry of the conformational changes [2–7]. However, this formalism alone is insufficient to characterize the conformational motions. Thus, the formalism does not contain any physical criterion that would allow identifying the set of essential degrees of freedom for a particular protein. Ranking of the collective coordinates according to the associated mean-square displacements only compares the displacements relative to each other and does not identify, what is the "sufficient" value of the displacement for a coordinate to be essential. A physical criterion for distinguishing the essential motion still needs to be derived. Furthermore, protein isoforms sometimes show only minor geometrical differences and yet have dramatically different functionalities. Composition of solvent is another factor, whose impact is difficult to capture by analyzing the geometry of the phase trajectory alone. Thus, the dynamics of the collective motions needs to be addressed in addition to their characterization through statistical techniques.

III. DYNAMICS OF CONFORMATIONAL MOTIONS IN MACROMOLECULES

The incremental effort to create a theory of conformational dynamics in proteins is represented, for example, by Refs. [4,9,13–21] and citations therein. Thus, it has been suggested to describe dynamics of proteins by the classic Langevin equations of motion, with the dynamic variables represented by either the essential collective coordinates x [4,19] or by the Cartesian coordinates of atoms X [14]. Accordingly, significant effort has been invested into evaluation of the potentials of mean force [9,19] and friction or diffusion coefficients [4,13,15,16,19] that define the dynamics of

a protein molecule embedded in a solvent. Several authors proposed employing the generalized Langevin equation as a more comprehensive model for proteins [15,17,18]. In the recent study [21], an analogy with the Mori projection operator formalism [11] has been employed to postulate that motion along the collective coordinates can be described by the generalized Langevin equation as well. Based on this assumption, an approach has been developed that represents protein dynamics by motion along a single collective coordinate that has been derived through the PCA technique [21]. However, applicability of the generalized Langevin equation to the essential degrees of freedom extracted from molecular-dynamic trajectories has not been proven rigorously. In particular, the relation between the Langevin equation for Cartesian coordinates of atoms in the protein and those for the collective essential coordinated derived through PCA have not been established. Also the restriction of the theory to a single collective degree of freedom is too a crude approximation for realistic proteins. Below a general *ab initio* formalism is developed to characterize dynamics of proteins based on the multivariate analysis of their atomistic trajectories.

A. Equations of motion for the projected Cartesian coordinates of atoms $X_i^E(t)$

Consider a system composed of a protein embedded in a solvent and containing N atoms. The phase trajectory of the entire system is given by the vector $\vec{X}(t) = \{X_1(t), X_2(t), \ldots, X_{3N}(t)\}$, which is a function of time. The Cartesian coordinates of atoms, X_i, obey the equations of motion, $\ddot{X}_i = m_i^{-1} F_i$, where m_i are the masses of atoms and F_i are the forces acting along the coordinates X_i. This system of $3N$ equations can be represented by

$$\ddot{\vec{X}} = m^{-1}\vec{F}(\vec{X}), \tag{10}$$

where m is a diagonal matrix providing the masses of atoms. Employing Eq. (6), one can write $\vec{F}(\vec{X}) = \vec{F}(\vec{X}^E + \vec{X}^{1-E})$, where $\vec{X}^E$ and $\vec{X}^{1-E}$ are, respectively, the projections of the phase trajectory onto the subspace of essential degrees of freedom and onto that of the fluctuations, as explained in Sec. II. By definition, $\vec{X}^{1-E}$ represents minor changes in atoms' positions as compared to a more pronounced essential motion given by $\vec{X}^E$. Accordingly, one can use the Taylor expansion for the force in Eq. (10),

$$\begin{aligned}\vec{F}(\vec{X}) &= \vec{F}(\vec{X}^E + \vec{X}^{1-E}) \\ &\approx \vec{F}(\vec{X}^E) + \frac{\partial \vec{F}(\vec{X}^E)}{\partial \vec{X}^E}\vec{X}^{1-E} \\ &= \vec{F}(\vec{X}^E) - K\vec{X}^{1-E}.\end{aligned} \tag{11}$$

Here, $\vec{F}(\vec{X}^E)$ is the mean force, $K\vec{X}^{1-E}$ represents fluctuations of the force, and K is the matrix with the elements $K_{ij} = -\frac{\partial F_i}{\partial X_j}$. The exact equation of motion (10) can thus be replaced with the approximation

$$\ddot{\vec{X}} = m^{-1}[\vec{F}(\vec{X}^E) - K\vec{X}^{1-E}]. \tag{12}$$

Next, the projection operators P and $1-P$ defined by Eq. (7) are applied to both sides of Eq. (12), which gives the equations of motion for $\vec{X}^E$ and $\vec{X}^{1-E}$, respectively:

$$\ddot{\vec{X}}^E = Pm^{-1}[\vec{F}(\vec{X}^E) - K\vec{X}^{1-E}], \tag{13}$$

$$\ddot{\vec{X}}^{1-E} = (1-P)m^{-1}[\vec{F}(\vec{X}^E) - K\vec{X}^{1-E}], \tag{14}$$

where Eq. (9) has been taken into account. The new operators Pm^{-1} and $(1-P)m^{-1}$ that appear in the right-hand sides of Eqs. (13) and (14) represent the mass-weighted projections onto the subspaces of the essential motions and fluctuations, respectively.

To proceed further an assumption is required, that the coordinates $\vec{X}^{1-E}$ change significantly faster that $\vec{X}^E$, so that the elements of the matrix K in Eq. (14) can be considered as constants (this implies that the fluctuations can also be considered as an ensemble of harmonic high-frequency oscillations). This makes Eq. (14) solvable with respect to $\vec{X}^{1-E}$, for example, through the Laplace transform technique. Skipping standard but cumbersome intermediate steps, the general solution of Eq. (14) is given by

$$\begin{aligned}\vec{X}^{1-E}(t) = &\int_0^t Z(t-\tau)(1-P)m^{-1}\vec{F}(\vec{X}^E(\tau))d\tau \\ &+ \vec{R}(t, \vec{X}^{1-E}(0), \dot{\vec{X}}^{1-E}(0)).\end{aligned} \tag{15}$$

The first term in right-hand side of Eq. (15) has the form of the memory integral with the damping kernel $Z(t)$, which in a general case is a nondiagonal matrix [22]. The expression $(1-P)m^{-1}\vec{F}(\vec{X}^E)$ under the integral represents the mass-weighted projection of the force $\vec{F}(\vec{X}^E)$ acting within the subspace of essential motions onto the subspace of fluctuations. This can be rephrased as coupling of the fluctuations with the essential degrees of freedom. The second term in the right-hand side of Eq. (14), which is symbolically represented by the vector $\vec{R}$, is a linear combination of harmonic functions of time [26] weighted with the values of $\vec{X}^{1-E}(0)$ and $\dot{\vec{X}}^{1-E}(0)$ at the initial time $t=0$. $\vec{R}$ can be viewed as the contribution of random noise to $\vec{X}^{1-E}$.

In the case when the harmonic oscillations representing the fluctuations are coupled bilinearly with a slower and conceivably anharmonic essential motion [23–29], the solution of Eq. (14) is

$$\vec{X}^{1-E}(t) = \int_0^t \dot{Z}_H(t-\tau)\vec{X}^E(\tau)d\tau + \vec{R}_H(t), \tag{16}$$

which can be rewritten, equivalently,

$$\vec{X}^{1-E}(t) = Z_H(0)\vec{X}^E(t) - Z_H(t)\vec{X}^E(0) - \int_0^t Z_H(t-\tau)\dot{\vec{X}}^E(\tau)d\tau + \vec{R}_H(t). \quad (17)$$

This approximation is known in the literature as the model of harmonic bath, or bath of harmonic oscillators [25–29], and has been suggested as a reasonable approach to handle the fluctuations in macromolecules [25,28]. Details regarding the form of the damping kernel $Z_H(t)$ and the random function $\vec{R}_H(t)$ for the model of harmonic bath can be found, for example, in Refs. [24,26]. Substitution of Eq. (17) into the right-hand side of Eq. (13) provides the equation of motion for the projected coordinates $\vec{X}^E(t)$,

$$\ddot{\vec{X}}^E = Pm^{-1}\left[\vec{F}(\vec{X}^E) + KZ_H(0)\vec{X}^E(t) - K\int_0^t Z_H(t-\tau)\dot{\vec{X}}^E(\tau)d\tau - KZ_H(t)\vec{X}^E(0) + K\vec{R}_H(t)\right], \quad (18)$$

which can be converted into a form that resembles the generalized Langevin equation,

$$\ddot{\vec{X}}^E = Pm^{-1}\left[-\frac{\partial U(\vec{X}^E)}{\partial \vec{X}^E} - \int_0^t Z_X(t-\tau)\dot{\vec{X}}^E(\tau)d\tau + \vec{R}_X(t)\right], \quad (19)$$

where

$$\frac{\partial U(\vec{X}^E)}{\partial \vec{X}^E} = -\vec{F}(\vec{X}^E) - KZ_H(0)\vec{X}^E(t),$$

$$Z_X(t) = KZ_H(t),$$

$$\vec{R}_X(t) = -KZ_H(t)\vec{X}^E(0) + K\vec{R}_H(t). \quad (20)$$

Here $U(\vec{X}^E)$ is the potential of mean force, the expression $-\int_o^t Z_X(t-\tau)\dot{\vec{X}}^E(\tau)d\tau$ is the dissipative force with the memory kernel $Z_X(t)$, and $\vec{R}_X(t)$ can interpreted as the random force, in the sense that $\vec{R}_X(t)$ does not depend on the dynamics of the system considered [23] and satisfies the requirements $\langle R_{X,i}(t)\rangle = 0$ and $\langle R_{X,i}(0)R_{X,j}(t)\rangle = \beta^{-1}Z_{X,ij}(t)$ [17,26]. The final step is rewriting of Eq. (19) in the form of the set of equations of motion for $3N$ atomic coordinates X_i^E:

$$\ddot{X}_i^E(t) = -\sum_{j=1}^{3N} M_{ij}^{-1}\frac{\partial U}{\partial X_j^E} - \sum_{l=1}^{3N}\int_0^t \Xi_{il}(t-\tau)\dot{X}_l^E(\tau)d\tau + \rho_i(t), \quad i = 1,2,\ldots,3N. \quad (21)$$

Here

$$M_{ij}^{-1} = \sum_{k=1}^{k_{\max}} E_i^k E_j^k m_j^{-1}, \quad (22)$$

$$\Xi_{il}(t) = \sum_{j=1}^{3N} M_{ij}^{-1} Z_{X,lj}(t), \quad (23)$$

$$\rho_i(t) = \sum_{j=1}^{3N} M_{ij}^{-1} R_{X,j}(t). \quad (24)$$

The system of equations of motion (21) describes trajectories for all atoms in the system, projected onto the subspace of essential degrees of freedom. The first term on the right-hand side represents a "purely" essential motion defined by the mean force $-\partial U/\partial X_j^E$ and by the effective mass M_{ij}. The other terms on the right-hand side describe the influence of fluctuations onto the essential motion. The fluctuations manifest themselves in the form of the dissipative force $-\int_0^t \Xi_{il}(t-\tau)\dot{X}_l^E(\tau)d\tau$ and the random force $\rho_i(t)$. It is noteworthy that the $3N$ atomic degrees of freedom are coupled through the summations in the right-hand side of Eq. (21).

B. Equations of motion for the essential collective coordinates $x^k(t)$

The system of generalized Langevin equations (21) describes the trajectories of all N atoms in the protein (and also in the solvent around it). Although the trajectories have been projected to be representative of essential motions in the system, still $3N$ coupled Cartesian coordinates are involved in the theory. To build efficient coarse-grained models, the number of coordinates must be decreased. This can be reached if the collective coordinates x^k are considered instead of the Cartesian coordinates of individual atoms X^E [see Eq. (3) for the definition of x^k]. Indeed, the collective coordinates can be ranked according to the respective mean-square displacements and the truncated set of collective coordinates can be considered as explained in Sec. II. By the analogy with the theory of chemical kinetics [29], it has been suggested in the literature to consider the conformation changes in a protein as a generalized chemical reaction and view the functions $x^k(t)$ as the generalized reaction coordinates that represent the conformation changes [4,9,10]. Each of these generalized conformational coordinates represents one of the collective degrees of freedom in the system. Below the equations of motion are derived for a set of essential collective coordinates, $x^1, x^2, \ldots, x^{k_{\max}}$. The number of the essential coordinates, $k_{\max}$, is not determined or limited at this point; however, it is assumed that $k_{\max} \ll 3N$. This is consistent with the reported analysis of mean-square displacements in peptide molecules [2–7,9,28], which shows that in most cases, only five to ten essential collective coordinates are responsible for 70%–90% of the total mean-square displacement.

The essential collective coordinates x^k can be obtained through the following projection:

$$(\vec{X}^E \cdot \vec{E}^k) = \sum_{i=1}^{3N} E_i^k X_i^E = x^k, \quad k = 1,2,\ldots,k_{\max}. \quad (25)$$

The equations of motion for $x^k(t)$ are provided by a similar projection applied to both sides of Eq. (19),

$$(\ddot{\vec{X}}^E \cdot \vec{E}^k) = \ddot{x}^k = \left(\vec{E}^k \cdot Pm^{-1} \left[-\frac{\partial U(\vec{X}^E)}{\partial \vec{X}^E} - \int_0^t Z_X(t-\tau)\dot{\vec{X}}^E(\tau)d\tau + \vec{R}_X(t) \right] \right). \quad (26)$$

To convert Eq. (26) into a more convenient form, some modifications are required. First, let us note that for an arbitrary vector $\vec{Y}$, $(\vec{E}^k \cdot P\vec{Y}) = \Sigma_{l=1}^{k_{\max}} (\vec{E}^k \cdot \vec{E}^l)(\vec{E}^l \cdot \vec{Y}) = (\vec{E}^k \cdot \vec{Y})$, and thus the operator P on the right-hand side of Eq. (22) can be omitted. Second, by definition, $\vec{X}^E = \Sigma_{k=1}^{k_{\max}} \vec{E}^k x^k$, which leads to the change of variables $\frac{\partial U}{\partial \vec{X}^E} = \Sigma_{k=1}^{k_{\max}} \vec{E}^k \frac{\partial U}{\partial x^k}$. Third, it is convenient to introduce the vectors $\vec{x}^k = \vec{E}^k x^k$ and $\vec{f}^k = -\vec{E}^k \frac{\partial U}{\partial x^k}$, which represent, respectively, the kth essential collective coordinate and the mean force associated with it. With these improvements, the equation of motion for x^k becomes

$$\ddot{x}^k = \sum_{l=1}^{k_{\max}} (\vec{E}^k \cdot m^{-1} \vec{f}^l) - \sum_{l=1}^{k_{\max}} \int_0^t [\vec{E}^k \cdot m^{-1} Z_X(t-\tau)\dot{\vec{x}}^l(\tau)]d\tau + [\vec{E}^k \cdot m^{-1} \vec{R}_X(t)]. \quad (27)$$

The right-hand side of Eq. (27) can also be represented in the scalar form

$$\ddot{x}^k = -\sum_{l=1}^{k_{\max}} {\mu_{kl}}^{-1} \frac{\partial U}{\partial x^l} - \sum_{l=1}^{k_{\max}} \int_0^t \xi_{kl}(t-\tau)\dot{x}^l(\tau)d\tau + r^k(t), \quad k = 1,2,\ldots,k_{\max}, \quad (28)$$

where

$${\mu_{kl}}^{-1} = \sum_{i=1}^{3N} E_i^k E_i^l m_i^{-1}, \quad (29)$$

$$\frac{\partial U}{\partial x^k} = \sum_{i=1}^{3N} E_i^k \frac{\partial U}{\partial X_i^E}, \quad (30)$$

$$\xi_{kl}(t) = \sum_{i,j=1}^{3N} E_i^k E_j^l {m_j}^{-1} Z_{X,ij}(t), \quad (31)$$

$$r^k(t) = \sum_{i=1}^{3N} E_i^k m_i^{-1} R_{X,i}(t). \quad (32)$$

Equation (28) describes conformational motions in a protein in terms of statistically independent collective coordinates x^k. However, from Eq. (28) it is evident that any essential collective coordinate x^k is coupled dynamically to all other essential collective coordinates in the system. Coupling of the terms representing the mean forces $-\mu_{kl}^{-1} \partial U / \partial x^k$ is provided by the matrix of effective mass μ_{kl}^{-1}, and coupling of the dissipative forces $-\int \xi_{kl}(t-\tau)\dot{x}^l(\tau)d\tau$ is provided by the memory matrix $\xi_{kl}(t)$. In a general case, both matrices are nondiagonal and thus coupling is present. However, two cases exist when the equation of motion (28) can be simplified. In the first case, all atoms in a hypothetic system have the same mass m. In practice, this is encountered when only trajectories of C_α atoms are considered. Then, the matrix μ_{kl}^{-1} becomes diagonal,

$${\mu_{kl}}^{-1}|_{k=l} = m^{-1},$$

$${\mu_{kl}}^{-1}|_{k\neq l} = 0, \quad (33)$$

and the first term on the right-hand side of Eq. (28) reduces to simply $-m^{-1} \partial U / \partial x^k$; e.g., coupling of the mean forces is eliminated. Coupling of the dissipative forces remains, however. Another, even simpler case occurs when a single essential coordinate is sufficient to describe the system ($k_{\max} = 1$) [21]. Only under this assumption is the equation of motion (28) confined to a single collective degree of freedom. Note that in this case the effective mass is given by

$$\mu^{-1} = \sum_{i=1}^{3N} (E_i^1)^2 {m_i}^{-1} \quad (34)$$

and can be interpreted as a weighted average of the masses of atoms, with the weights equal to $(E_i^1)^2$. Thus, the value $(E_i^1)^2$ represents a measure of involvement of the ith atomic degree of freedom in the collective dynamics of the system.

IV. EXAMPLE OF APPLICATION: IDENTIFICATION OF CORRELATED DYNAMIC DOMAINS AND DERIVATION OF EQUATIONS FOR THE DOMAIN DYNAMICS

One of the major and largely unsolved problems in the theory of protein dynamics is representation of the collective motion in terms of particular domains containing atoms that move in a coherent way. Despite a rather common expectation, the essential collective coordinates do not represent any particular groups of atoms explicitly [30]. Efforts at identification of dynamic domains based on molecular simulations of proteins have been recently reviewed in Ref. [31]. Difficulties arise even with the very definition of domains, which vary from paper to paper, and sometimes include rather vague criteria such as being a visually recognizable substructure in the protein [31]. In the most elaborate approach [32–36] dynamic domains are defined as rigid bodies and identified by clustering of translations and rotations of elementary building blocks. The problem of this approach, however, is that the elementary building blocks, such as residuals or groups of atoms, must be postulated *a priori*. Furthermore, the differences in motion that need to be captured are very subtle and susceptible to uncertainties, such as the high-frequency fluctuations. In order to eliminate the random fluctuations, various filtering procedures are sometimes applied in parallel with the motions' clustering [35–37]. As a result, the methodologies of domain identification become computationally expensive and overwhelmingly complex. Results of domain identification depend on the assumptions made, particular techniques employed, sampling schemes applied at multiple intermediate steps, etc., which makes such results difficult to reproduce and interpret [31]. A universal and dynamically justified concept for identification of dynamic domains has not been suggested.

Based on the theory developed in this paper, a simple and physically transparent formalism of domain identification is introduced below, with examples for protein G. The approach does not employ any *a priori* assumptions regarding the structure of domains or their elementary building blocks and does not require any additional noise filtering since the domains are identified in the space of essential collective motions. Because the methodology of domain identification is based on a rigorous theoretical background, it can be employed as a starting point for further theoretical development. Thus, a system of generalized Langevin equations is derived in this work that describes motion of the correlated domains.

A. Definition of dynamic correlated domains

Consider the equations of motion for the projected trajectory (19). It is convenient to represent Eq. (19) in the form

$$\ddot{\vec{X}}^E = P\vec{Y}, \tag{35}$$

where

$$\vec{Y} = m^{-1}\left[-\frac{\partial U(\vec{X}^E)}{\partial \vec{X}^E} - \int_0^t Z_X(t-\tau)\dot{\vec{X}}^E(\tau)d\tau + \vec{R}_X(t) \right].$$

According to Eq. (8), the projection operator P employs a number of mathematical operations involving the essential collective eigenvectors $\vec{E}^k$. These result in coupling of the projected atomic degrees of freedom through the matrices of effective mass M_{ij} and memory kernel $\Xi_{il}(t)$ [see Eqs. (21)–(23)]. It is therefore clear that any correlations in the system are implicitly present in the structure of the projection operation P, and these correlations can be identified through an analysis of the operator P as well as of the essential collective eigenvectors $\vec{E}^k$ that it employs.

To better understand, which kind of analysis is required to identify the correlations, let us consider the kth essential collective eigenvector, $\vec{E}^k=\{E_1^k, E_2^k, \ldots, E_{3N}^k\}$. By definition, the values E_i^k represent direction cosines of the vector $\vec{E}^k$ in the $3N$-dimensional phase space. The values E_i^k can also be viewed as the projections of the vector $\vec{E}^k$ onto Cartesian degrees of freedom of individual atoms. Since the system considered contains N atoms, the entire set of direction cosines $\{E_1^k, E_2^k, \ldots, E_{3N}^k\}$ can be subdivided into N subsets each containing three values $\{E_{n,x}^k, E_{n,y}^k, E_{n,z}^k\}$, where $n=1, \ldots, N$. Each of these subsets contains direction cosines relative x, y, and z Cartesian degrees of freedom of an individual atom. Each collective eigenvector can therefore be represented by

$$\vec{E}^k = \sum_{n=1}^{N} \vec{E}_n^k, \tag{36}$$

where $\vec{E}_n^k=\{E_{n,\alpha}^k\}$ and $\alpha=1$, 2, or 3 denotes the degrees of freedom x, y, and z. Employing Eq. (36) in the projection operator P on the right-hand side of Eq. (35) leads to

$$P\vec{Y} = \sum_{k=1}^{k_{\max}} \sum_{n_2=1}^{N} (\vec{E}_{n_2}^k \cdot \vec{Y}) \sum_{n_1=1}^{N} \vec{E}_{n_1}^k \tag{37}$$

or, in scalar form,

$$(P\vec{Y})_{n_1,\alpha} = \sum_{n_2=1}^{N} \sum_{\beta=1}^{3} C_{n_1,\alpha,n_2,\beta} Y_{n_2,\beta}, \tag{38}$$

where

$$C_{n_1,\alpha,n_2,\beta} = \sum_{k=1}^{k_{\max}} E_{n_1,\alpha}^k E_{n_2,\beta}^k. \tag{39}$$

Recall that the summations in the right-hand side of Eq. (38) represent coupling between particular atomic degrees of freedom.

The direction cosines $E_{n,\alpha}^k$ in Eq. (39) can adopt positive, negative, or zero values. In the first case, the collective mode represented by $\vec{E}^k$ is in phase with the atomic degree of freedom $\{n,\alpha\}$, in the second case it is in antiphase, and in the third case there is no correlation. From the discussion in Sec. III B it follows that the magnitude $|E_{n,\alpha}^k|$ is representative of the level of the correlation; the larger $|E_{n,\alpha}^k|$ is, the stronger is the involvement of the atomic degree of freedom $\{n,\alpha\}$ into the collective mode k. Since the direction cosines $E_{n,\alpha}^k$ represent correlations of collective degrees of freedom with individual atomic degrees of freedom, it is natural to define correlated domains as groups of atoms for which the values $E_{n,\alpha}^k$ have a similar magnitude for each of the essential collective degrees of freedom k [38]. Accordingly, the cross-correlation terms in Eqs. (38) and (39) can be classified into two categories.

(i) The atoms n_1 and n_2 belong to the same domain and their Cartesian degrees of freedom are similar ($\alpha=\beta$). In this case, the coefficients $C_{n_1,\alpha,n_2,\beta}$ given by Eq. (39) are nonzero and positive, because $E_{n_1,\alpha}^k=E_{n_2,\beta}^k$ for all k, and $E_{n_1,\alpha}^k \neq 0$ for at least some k [38]. Occurrences like this can be summarized by

$$(P\vec{Y})_{n_1,\alpha}^{(\mathrm{i})} = \sum_{n_2 \in \{N^\delta\}} C_{n_1,\alpha,n_2,\alpha} Y_{n_2,\alpha}. \tag{40}$$

Here, δ denotes the domain, $\{N^\delta\}$ denotes the set of atoms in the domain δ, and the expression $n_2 \in \{N^\delta\}$ means that the atom n_2 belongs to the domain δ.

(ii) The atom n_1 belongs to a correlated domain δ, but the atom n_2 does not belong to this domain, $n_2 \notin \{N^\delta\}$, and/or the Cartesian degrees of freedom are different, $\alpha \neq \beta$. The contributions of such cases are represented by the following expression:

$$(P\vec{Y})_{n_1,\alpha}^{(\mathrm{ii})} = \sum_{n_2=1}^{N} \sum_{\beta \neq \alpha} C_{n_1,\alpha,n_2,\beta} Y_{n_2,\beta} + \sum_{n_2 \notin \{N^\delta\}} C_{n_1,\alpha,n_2,\alpha} Y_{n_2,\alpha}. \tag{41}$$

Now, the values $E_{n_1,\alpha}^k$ and $E_{n_2,\beta}^k$ in Eq. (39) can differ in both magnitude and sign. Accordingly, the coefficients $C_{n_1,\alpha,n_2,\beta}$ are given by a summation of both positive and negative terms, which generates smaller values of $|C_{n_1,\alpha,n_2,\beta}|$ than in the case (i).

The entire right-hand side of the equation of motion for an atom in a correlated domain reads

$$(P\vec{Y})_{n_1,\alpha} = (P\vec{Y})^{(\mathrm{i})}_{n_1,\alpha} + (P\vec{Y})^{(\mathrm{ii})}_{n_1,\alpha}. \quad (42)$$

The first contribution in Eq. (42) describes coupling of degrees of freedom which are correlated within a domain, whereas the second contribution corresponds to coupling of degrees of freedom which do not show a strong correlation within the same domain. From the previous discussion it follows that for domains that weakly interact with each other, the first contribution is larger than the second one, $|(P\vec{Y})^{(\mathrm{i})}_{n_1,\alpha}| > |(P\vec{Y})^{(\mathrm{ii})}_{n_1,\alpha}|$. In particular cases one can expect $|(P\vec{Y})^{(\mathrm{i})}_{n_1,\alpha}| \gg |(P\vec{Y})^{(\mathrm{ii})}_{n_1,\alpha}|$, so that Eq. (42) can be approximated by the first contribution alone, $(P\vec{Y})_{n_1,\alpha} \approx (P\vec{Y})^{(i)}_{n_1,\alpha}$.

This result demonstrates the physical meaning of domains in the present theory. Domains are groups of atoms that show a strong dynamic coupling in the generalized Langevin equation of motion. The domains are identified as groups of atoms, for which the direction cosines of the essential collective degrees of freedom $E^k_{n,\alpha}$ adopt similar values for each k. Unlike existing approaches to identifying domains in proteins, in this work (i) subject to clustering are the directional cosines of the essential collective degrees of freedom, *and not* translations and/or rotations of individual atomic groups in the Cartesian space, which makes the formalism generically immune to fluctuations; (ii) no assumptions regarding any elementary building blocks and/or interatomic interactions are made, and therefore the formalism is universal and applicable to any kind of systems; and (iii) the identification of domains is intimately related to the essential dynamics of proteins, which makes the formalism simple and physically transparent, and its outcomes are easy to interpret. An example of the domain identification with interpretations is given below.

B. Example of correlated domains in protein G

This section gives an example of correlated dynamic domains in the fragment B1 of protein G. The PDB structure 1igd was used for this purpose. A molecular dynamics trajectory of the solvated protein was generated using the GROMACS 3.2.1 code with the GROMOS96 force field [39]. After equilibration, 2000 snapshots were taken every 0.1 ps. This 0.2-ns trajectory was processed by PCA taking into account all atoms in the protein. For essential collective coordinates, 10 principal components with the highest eigenvalues were used for this example ($k_{\max}=10$) [40]. The direction cosines of the collective coordinates $E^k_{n,\alpha}$ have been represented by N points (N is equal to the total number of atoms), each corresponding to an individual atom, in the $3k_{\max}$-dimensional space of essential collective motions. In this space, points that are located close to each other represent a strong correlation in motion of corresponding atoms. To obtain dynamic domains, the N points have been clustered using the nearest-neighbor technique [41]. The advantage of this technique it that no structural property, such as, e.g., the number of domains, needs to be postulated. A potential challenge, however, is that the interdomain distance d needs to be identified, which defines the maximum distance in the $3k_{\max}$-dimensional space, for the corresponding atoms to belong to the same domain [42]. The identification of the domains is sensitive to the selection of d, and the consistent magnitude of d depends on the dynamics of a particular protein. Thus, Fig. 1(a) shows a dependence of the number of domains in protein G as a function of the distance d, and Fig. 1(b) shows the total number of atoms that are included in domains as a function of d [43]. From the figures it is evident that a minimum distance $d_{\min}$ can be identified, beyond which no correlated domains can be found. In Fig. 1(a) this distance is close to 0.002. A maximum interdomain distance $d_{\max} \approx 0.0065$ is also visible, for which most of the protein molecule is recognized as a single domain. This is demonstrated also by Figs. 1(b) and 1(c), which show the total number of atoms in all domains, N_{tot}, the number of atoms in the largest domain, N_1, and their difference $N_{tot}-N_1$ as functions of d. From Fig. 1(c) it is possible to identify the value $d \approx 0.0038-0.0040$, which maximizes the difference $N_{tot}-N_1$ and therefore provides the most informative breakdown of the protein into domains. Note that the dependence of the total number of atoms involved in domains in Fig. 1(b) shows a trend to saturate at distances $d \approx 0.0038-0.0042$, which is close to the optimum distance d defined from Fig. 1(c). The impact of the interdomain distance is addressed further in Sec. IV D, whereas here, the optimum interdomain distance $d=0.0039$ is adopted as a consistent condition for clustering [43].

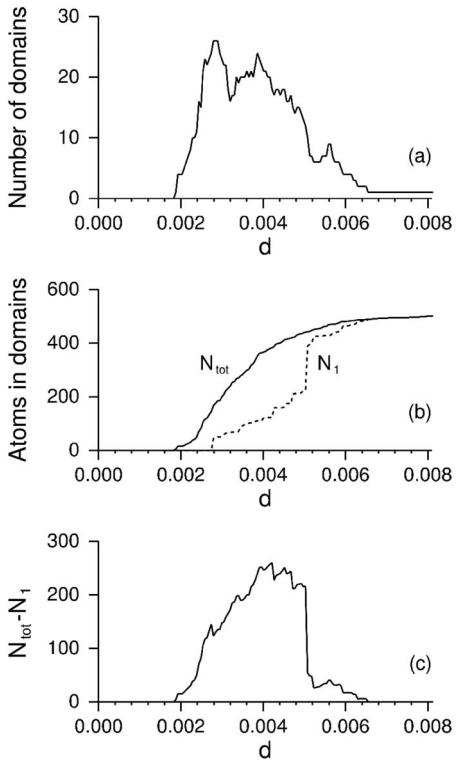

FIG. 1. The number of domains in fragment B1 of protein G containing more than two atoms, as a function of the interdomain distance d (a); the number of atoms in all domains, N_{tot} (solid line), and the number of atoms in the largest domain, N_1 (dashed line), as functions of d (b); and the difference $N_{\mathrm{tot}}-N_1$ (c) [42].

Figure 2(a) shows three largest dynamic domains that have been identified in protein G. In the figure, the domains

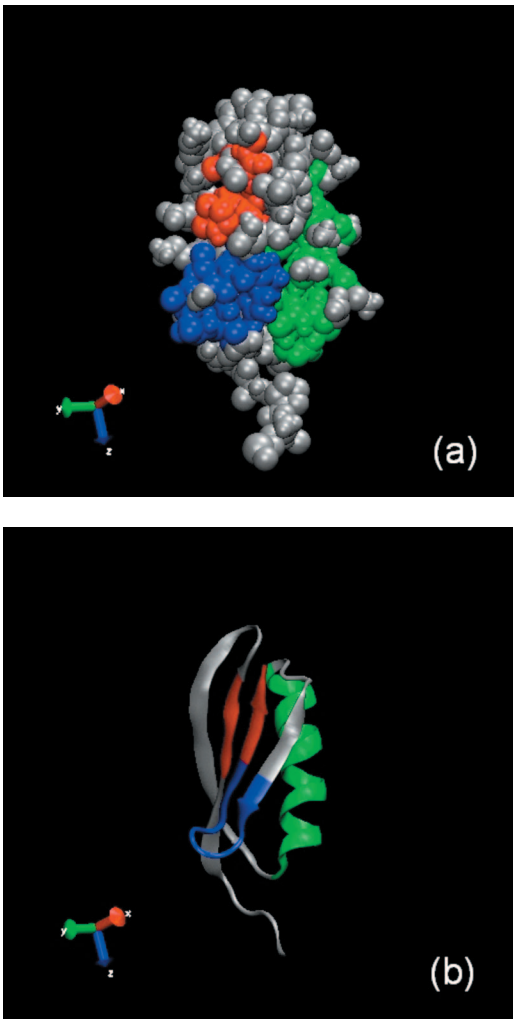

FIG. 2. (Color) Three largest domains identified in protein G for $k_{max}=10$: general view (a) and schematic (b). The domains are shown with green, red, and blue.

are shown with green, red, and blue colors, whereas atoms that do not belong to the three largest domains are colored gray. It can be seen, first of all, that the domains form compact groups of atoms. This is an interesting and promising result, since the clustering formalism employed in this paper does not require any proximity of atoms' locations in the primary, secondary, or tertiary structure. By the definition, a proximity in the $3k_{max}$-dimensional space of essential collective motions reveals only directional correlations in motion. The fact that these correlations identify compact atomic groups confirms the viability of the clustering formalism. Another noticeable feature is that some side groups connected to the correlated clusters have not been recognized as belonging to these clusters. Understandably, the side groups have less spatial constraints and more flexibility in comparison to the main-chain groups, which results in a weaker correlation.

Table I compares the residue sequences of the domains with the secondary structure in protein G, and Fig. 2(b) shows the corresponding schematic sketch. From both the table and the figure, it can be seen that some sequences of residues in the domains follow both the primary and secondary structure in the protein. Thus, the first domain in Table I, which is colored green in Fig. 2, follows closely the α-helix. At the same time, the second domain (colored red) contains a part of the first β-hairpin as well as a part of the second β-hairpin, which are adjacent to each other in the tertiary structure, but quite remotely separated in the main chain. The third domain (colored blue) contains the $\beta3$-$\beta4$ loop in the second β-hairpin.

TABLE I. Comparison of domains identified in protein G with the secondary structure.

Three largest domains identified in this work for $k_{max}=10$		Secondary structure	
Domain[a]	Residues	Element	Residues
1	26–41	α-helix	28–42
2	10–13; 58–61	First β-hairpin: $\beta1$- and $\beta2$-strands	6–25
3	50–57	Second β-hairpin: $\beta3$- and $\beta4$-strands	47–60

[a]The domains labeled by 1, 2, and 3 are shown in Fig. 2 with green, red, and blue, respectively.

To conclude, the present theory leads to a dynamically consistent definition of correlated domains in proteins. In the example considered, the identified domains are composed of compact groups of atoms. The domains have also shown a reasonable match with the secondary structure; however, there is no complete similarity. Some domains can contain entire elements of secondary structure (with the exception for flexible endgroups), other domains include only parts of such elements, and still others are composed of different elements that are located near each other in the tertiary structure.

C. Role of the number of essential coordinates k_{max}

In the conventional PCA of molecular dynamic trajectories, the number of essential coordinates k_{max} is not a well-established quantity, so that an appropriate k_{max} is defined in each case individually [44]. In the present theory, the number of essential coordinates affects the structure of the projection operator P [see Eq. (8) and Sec. IV A]. Perhaps the most simple and obvious outcome of this dependence is the dimensionality $3k_{max}$ of the space of directional cosines E_i^k, where correlated dynamic domains are identified. The sketch in Fig. 3 clarifies the impact of this dimensionality on the identification of dynamic domains. The plane $\{E_1$-$E_2\}$ in Fig. 3 represents a "high-dimensional" essential space, the axis E_1 represents a "low-dimensional" subspace, and the points represent sets of the directional cosines for individual atoms. As can be seen in Fig. 3, a high-dimensional essential space reveals differences in positions of the points that cannot be captured in a low-dimensional subspace. At the same time, clustering in the subspace is apt to reveal subtle features of the projection into this subspace, not all of which are captured through clustering with a higher dimensionality. As a result, there is no direct correspondence between the outcomes of clustering with different values of k_{max}. Thus, at-

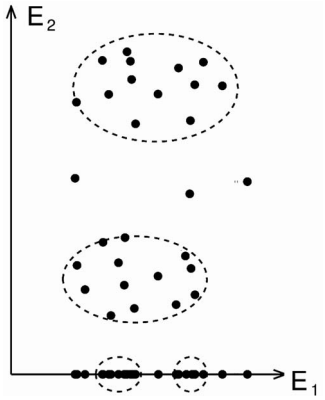

FIG. 3. A sketch illustrating identification of domains with different numbers of essential coordinates k_{max}. The plane $\{E_1$-$E_2\}$ represents a "high-dimensional" essential space, and the axis E_1 represents a "low-dimensional" subspace. The dashed lines indicate the hypothetic domains identified with the different dimensionalities.

oms that do not belong to any domain in a space k_{max1} can be recognized as a part of a domain in its subspace $k_{max2} < k_{max1}$ and vice versa.

As an example, Fig. 4 shows three largest domains identified in protein G with $k_{max}=1$, 2, 5, 10, 20, and 40 [45]. When only one essential coordinate is used, $k_{max}=1$, does the domain structure show a long-range network traversing the entire four-strand β-sheet. When k_{max} increases, correlations along the individual β-strands become more pronounced. At the same time, correlations across the β-sheet become limited to three β-strands with $k_{max}=2$ and include only pairs of β-strands with $k_{max}=5$ and higher. After k_{max} reaches 10, a further increase in k_{max} does not lead to any significant changes in the domain system. Consistently reproduced with increasing k_{max} are (i) the domain containing the α-helix (colored green in Fig. 4), (ii) the domain containing a part of the β1-strand (shown in red), and (iii) the domain containing the β3-β4 loop (shown in blue). The only exception is the region of the β4-strand between residues 58 and 61, which is identified either in or out the second domain when k_{max} changes. Evidently, the position at the *C*-terminus makes this region moderately flexible, which results in the variability of its identification. This said, identification of domains is largely robust when k_{max} increases beyond 10. It worth noting that the dimensionality of the essential space $k_{max}=10$ at which the robustness is reached corresponds to approximately 90% of the total displacement. The fact that robustness is observed for the essential dimensionalities that sample 90% of the total displacement or more is in lines with the basic assumption of this theory. Indeed, when doing the Taylor expansion in Eq. (11), it has been assumed that the displacement related to the essential motions, $\vec{X}^E$, is significantly larger than the fluctuations, which is compatible with the condition $|\vec{X}^E|/|\vec{X}| \approx 0.9$ at which robustness is reached.

A further discussion of the structural trends seen in Fig. 4 is given in the next section. At this point, the conclusion is that identification of domains with various numbers of essential coordinates, k_{max}, reveals complementary aspects of the

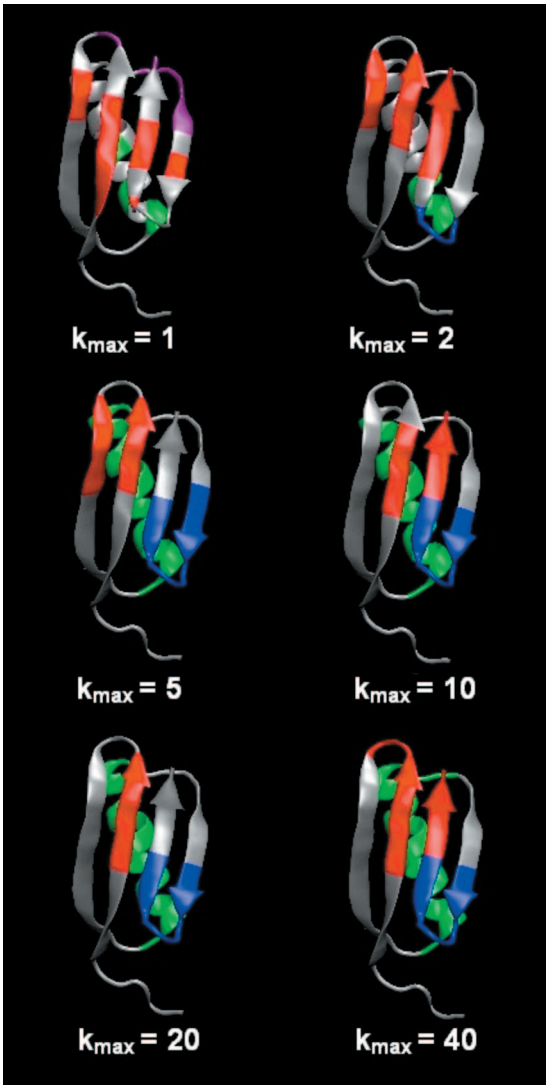

FIG. 4. (Color) Comparison of three largest domains identified in protein G for $k_{max}=1$, 2, 5, 10, 20, and 40 [45].

protein's dynamics. Clustering in low-dimensional subspaces reveals the structure of the averaged motion in the corresponding projections, whereas using large k_{max} provides a more complete accounting for the motion variability over multiple essential dimensions. Robustness of the domain identification is reached only with sufficiently large k_{max}, which in the case of protein G is close to 10. This is consistent with the basic requirement of the present theory, for the essential motion encompasses a major portion of the total displacement to provide a compatible representation of protein dynamics.

D. Comparison with NMR experiments

The dynamics of proteins has been extensively studied by nuclear magnetic resonance (NMR) methods [46–50]. Thus, numerous published NMR results for protein G [49–56], in principle, contain a rich pool of various dynamic information. However, explicit quantitative interpretation of NMR experiments in terms of protein's structure is an extremely

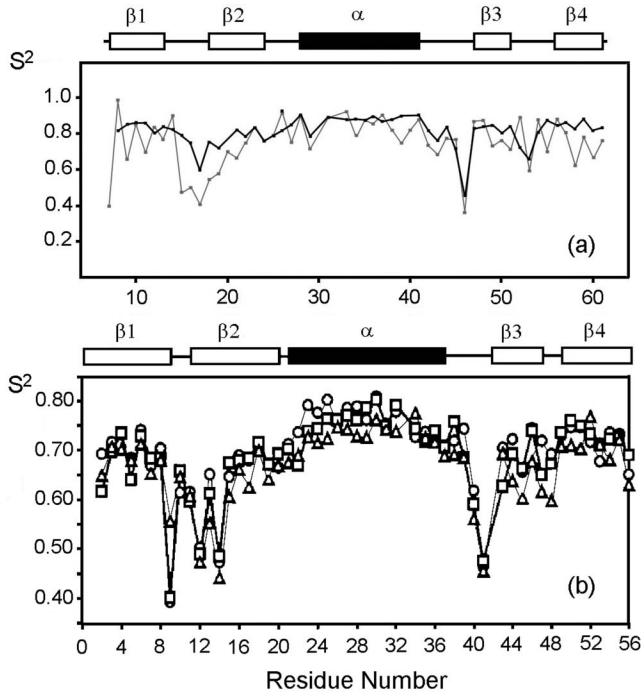

FIG. 5. Experimental NMR-derived order parameters S^2 for streptococcal protein G [57]: (a) adapted from Fig. 7 of Ref. [55] and reprinted in part with permission of The National Academy of Sciences of the USA; (b) adapted from Fig. 1(a) of Ref. [52] and reprinted in part with permission of the American Chemical Society. In (a), the bold line shows the relaxation-derived order parameter in the subnanosecond time scale [54] and thin line shows the order parameter derived from RDC data in the submillisecond scale [55]. In (b), relaxation data were employed representing the subnanosecond regimes [52]. The circles, squares, and triangles correspond to different mutants introduced to minimize cross-strand interactions [52].

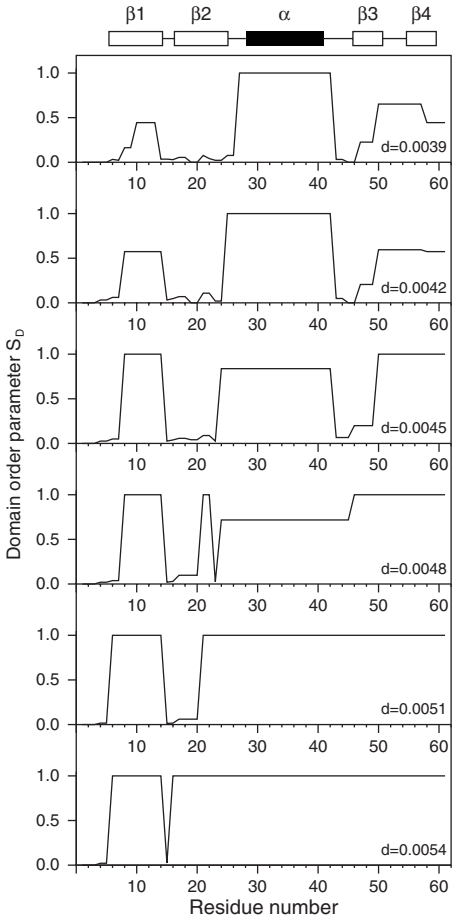

FIG. 6. The domain order parameters S_D computed for $k_{\max}=10$ and progressively increasing interdomain distance d. See Fig. 7 for the largest domains identified in each case.

challenging task for a number of reasons of both technical and fundamental origins [46–50]. This section gives an example, how comparison of the theory with NMR experiments can be employed to interpret some of the observations.

Perhaps the most often reported and well-reproduced experimental dynamic characteristic for protein G is the generalized order parameter S^2 [47,48], which represents the relative flexibilities of the protein's backbone locally for each residue. A number of experimental methodologies to obtain S^2 exist, which differ in the time scale addressed, kind of NMR data employed to derive S^2, model assumptions made, etc. Considered here are the most general properties of the order parameter S^2, which have been repeatedly observed in various studies. Most of these trends can be seen in Fig. 5, which shows examples of the S^2 profiles in protein G obtained by three independent research groups [52,54,55] through various techniques (see also caption to Fig. 5) [57]. It can be seen that consistently reproduced are high S^2 values in the regions of the α-helix, β1-strand, and β4-strand, which indicates that these regions are relatively rigid. In contrast, the loop between the β1- and β2-strands, the adjacent region of the β2-strand, and the loop between the α-helix and the β3-strand are recognized as regions of relative softness with smaller parameters S^2.

It is believed that the order parameter S^2 is not interpretable directly in terms of dynamic domains [47,50]. Qualitative comparisons are possible, however. As an example, consider the theoretical order parameter $S_D=\tilde{M}^{\delta}/\tilde{M}^{\max}$, defined individually for each residue and equal to the ratio of the mass of the corresponding domain $\tilde{M}^{\delta}$ to the mass of the largest domain $\tilde{M}^{\max}$. Thus, all residues within the same domain have equal order parameters S_D by definition. In the largest domain all residues have $S_D=1$, whereas $S_D=0$ for residues that are not part of any domain. The domain order parameter S_D therefore indicates whether a residue is a part of a domain and how large this domain is. Figure 6 shows a series of plots of the theoretical parameter S_D in protein G obtained with $k_{\max}=10$. It should be particularly emphasized that the computed function S_D presented in Fig. 6 *is not* a theoretical analog of the experimental parameter S^2. The definition and meaning of the values S_D and S^2 are different and, therefore, a direct quantitative comparison of the dependencies in Figs. 5 and 6 is not possible. However, the qualitative trends exhibited by two dependences are comparable. Indeed, within sufficiently large domains, the backbone can be considered as relatively inflexible in motion. The corresponding residues will therefore have both high values of S_D (because they belong to large domains) and high order pa-

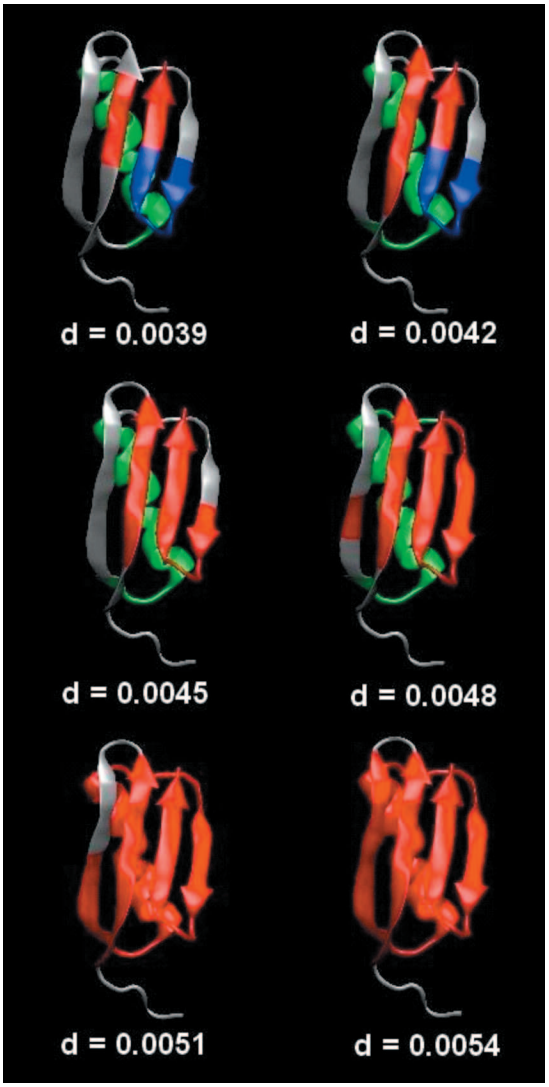

FIG. 7. (Color) Comparison of the largest domains in protein G for various interdomain distances d and $k_{max}=10$.

rameters S^2 (because large domains exhibit a relative rigidity). Otherwise, off-domain residues and those in small domains, which have low S_D values, should be less restricted in their motions, indicating the points of softness characterized by low S^2 as well. Thereby *relative* changes of the theoretic order parameter S_D as a function of the residue can be compared to similar *relative* changes of the experimental parameter S^2 [58].

It is convenient to analyze these structural trends by varying the interdomain distance d that affects the breakdown of the protein into dynamic domains as outlined in Sec. IV B. Figure 6 presents the theoretical order parameters S_D obtained for protein G with $k_{max}=10$ and various interdomain distances d, and Fig. 7 shows the corresponding domains. For the largest d values of 0.0054 and 0.0051, almost all protein is recognized as a single domain, with the exception of the N-terminus and the loop between the β1- and β2-strands. Accordingly, $S_D=1$ everywhere except for these regions, which thereby are recognized as the points of softness in agreement with the NMR experiments. When d decreases, a progressively increasing portion of the β2-strand is recognized as a region of softness. Residues 21 and 22, however, tend to ally with the rest of the β-sheet, which results in the oscillatory behavior of S_D in the region of the β2-strand for $d=0.0048$. For $d=0.0048$ and 0.0045, two separate domains are distinguished, one containing the α-helix and another containing most of the β-sheet, which results in a moderate decrease of the S_D level within the smaller domain. For $d=0.0045$ and less, the loop between the α-helix and the β3-strand, as well as the adjacent region of the β3-strand, acquires low S_D values, indicating a relative softness, which again agrees with the experiments. At $d=0.0042$, two different domains are recognized in the β-sheet, whose masses are considerably less than that of the domain containing the α-helix, which affects the shape of the profile. Nevertheless, the theoretical regions of the relative rigidity and softness show a clear qualitative resemblance with the experimental ones.

TABLE II. Comparison of predicted regions of relative softness (residue numbers) with experimental NMR S^2 data for protein G.

This theory	NMR data [52][a]	NMR data [54]	NMR data [55]
15–23	14–20	14–20	15–19
43–46	45–48	45–47	45–47

[a]The residue numbering system from Ref. [52] [Fig. 5(b)] has been amended to match that adopted in Fig. 5(a) and in the present theory.

The predicted regions of relative softness are compared with the experimentally determined ones in Table II. Here, the theoretical regions of maximum softness are determined from Fig. 6 considering the residues for which the value S_D is close to zero. The plot for $d=0.0045$ has been employed for this purpose, which corresponds to the maximum d value at which two major regions of softness are well recognized. The experimental regions of softness were determined from the major minima in the S^2 dependencies in Fig. 5. The close overlap of the theoretical and experimental regions of softness is obvious from Table II. This demonstrates that the major regions of relative softness emerging from the experimental S^2 profiles for protein G (obtained in various time regimes from subnanosecond [52,54] to submillisecond [55] scale) match very well those extracted from the domain system identified theoretically with the representative set of essential coordinates.

In addition to the conventional studies of the S^2 profiles, new methodologies for extracting complementary structural information from NMR experiments have undergone a rapid development [48,49]. Thus, in a recent study [55], analysis of residual dipolar coupling (RDC) was employed to obtain distributions of amplitudes of backbone motions in protein G. Some of these observed distributions exhibit a long-range networklike structure traversing the entire β-sheet, with adjacent β-strands showing similar motional amplitudes. These long-range cross-strand correlations, which have been interpreted as resulting from hydrogen bonding between adjacent β-strands [49,55], seemed to differ radically from the structure emerging from the conventional S^2 measurements.

In this work, a similar long-range cross-strand symmetry as reported in Ref. [55] is obtained by clustering with a single essential coordinate. As can be seen in Fig. 4 for $k_{\max}=1$, the domain colored red includes the adjacent regions in all four β-strands and the domain colored purple comprises a part of the β_1-β_2 loop, the C-terminus, and a part of the loop between the α-helix and the $\beta 3$-strand. Less involved in the long-range interstrand interactions are the regions shown in gray. These regions also traverse the entire β-sheet. The overall symmetry of the domain system is compatible with the directions of the interstrand hydrogen bonding, and the structure of the alternating cross-strand domain resembles closely that detected in Ref. [55] for the motional amplitudes.

Figure 4 shows however, that the extended domain structure traversing the entire β-sheet is obtained only with $k_{\max}=1$ and gradually disappears when the number of essential coordinates, $k_{\max}$, increases. This can be easily explained in light of the discussion of the role of the dimensionality $k_{\max}$ in Sec. IV C. The reason is that in protein G, the impact of the cross-strand bonding is superimposed over more pronounced interactions originating from the spatial constrains in the primary, secondary, and tertiary structures. These major impacts are fully captured only with a sufficient number of essential coordinates ($k_{\max}\sim 10$). The corresponding domain system accounts for all contributions; however, the impact of the cross-strand bonding is mostly overridden by stronger interactions. Otherwise, when only one essential coordinate is used for clustering, the complex multidimensional dynamics is not captured. Instead, the impact of interactions that have a simpler topology emerges. In the case of the β-sheet in protein G, clustering with $k_{\max}=1$ reveals correlations originating from the cross-strand bonding, which has a simple quasilinear geometry. Remarkably, a similar impact of the cross-strand interactions was observed experimentally in Ref. [55] through the RDC analysis, which is particularly sensitive to weak conformational fluctuations [56].

In conclusion, the domain system predicted for protein G with the dynamically representative set of essential coordinates, $k_{\max}=10$, identifies the major regions of rigidity and softness in agreement with the experimentally determined profiles of the order parameter S^2 [52,54,55]. At the same time, the domains obtained with only one essential coordinate match the RDC-based experimental distributions of backbone motions [55] that reflect the symmetry of the cross-strand hydrogen bonding.

The ability of the present approach to identify correlations with varying essential dimensionalities has a strong potential for interpretation of NMR data. The idea is to compare NMR results with the outcomes of the theoretic clustering and define the essential dimensionality at which the best match with experiment is achieved. As discussed above, the clustering with low dimensionalities reveals subtle features in the simplified (averaged) motion, whereas higher dimensionalities provide a more comprehensive and dynamically consistent description. Accordingly, NMR-derived dynamic data that can only be reproduced through the low-dimensional clustering provide insight into the most delicate details of protein structure; however, only those features that are identified theoretically with sufficiently high-dimensional essential spaces would be fully representative of the coarse-grained dynamics in the macromolecule.

E. Equations of motion for domains

A distinguishing feature and one of major advantages of the present identification of dynamic domains is that this approach is based on a rigorous theoretical background. Not only does this make the results transparent and easily interpretable, but the domain system identified is also available for further theoretical developments. In particular, after correlated domains are identified in a macromolecule, the coarse-grained dynamics can be described analytically by considering interaction of the domains with each other and with the environment. In this section, the formalism developed in Sec. III is applied to derive generalized Langevin equations for dynamic domains in a protein. The equations derived in this section are exact within the framework adopted in Sec. III. The only assumption made is that motion of a domain can be represented by its center of mass.

In the following discussion, each domain is characterized by the number of atoms involved N^δ, the domain's mass $\tilde{M}^\delta$, and the coordinates of the center of mass $\tilde{X}^\delta_\alpha$:

$$\tilde{M}^\delta = \sum_{n\in\{N^\delta\}} m_n, \tag{43}$$

$$\tilde{X}^\delta_\alpha(t) = \frac{1}{\tilde{M}^\delta}\sum_{n\in\{N^\delta_\alpha\}} m_n X^E_n(t). \tag{44}$$

Here δ denotes the domain, $\delta=1,2,\ldots,\delta_{\max}$, α denotes the x, y, or z coordinates, and X^E_i are coordinates of individual atoms. The expression $n\in\{N^\delta_\alpha\}$ on the right-hand side of Eq. (44) means that the summation is performed only for the αth coordinates of atoms involved in the domain δ.

The coordinates X^E_n can be expressed in terms of the essential collective coordinates x^k by $X^E_n=\Sigma_{l=1}^{k_{\max}}E^k_n x^k$, after which Eq. (44) takes the form

$$\tilde{X}^\delta_\alpha(t) = \sum_{k=1}^{k_{\max}} T_{\alpha\delta,k}x^k(t), \tag{45}$$

where $T_{\delta\alpha,k}=\frac{1}{\tilde{M}^\delta}\Sigma_{n\in\{N^\delta_\alpha\}}m_n E^k_n$. Double differentiation of Eq. (45) over time and replacing of $\ddot{x}^k$ with Eq. (28) gives

$$\ddot{\tilde{X}}^\xi_\alpha = -\sum_{k,l=1}^{k_{\max}} T_{\delta\alpha,k}\mu_{kl}^{-1}\frac{\partial U}{\partial x^l} - \sum_{k,l=1}^{k_{\max}} T_{\delta\alpha,k}\int \xi_{kl}(t-\tau)\dot{x}^l(\tau)d\tau + \sum_{k=1}^{k_{\max}} T_{\delta\alpha,k}r^k(t). \tag{46}$$

Equation (46) is a formal equation of motion for coarse-grained degrees of freedom represented by the coordinates of domains' centers of masses $\tilde{X}^\delta_\alpha$. If the total number of such coarse-grained degrees of freedom, $3\delta_{\max}$, is equal to the number of the collective coordinates, $k_{\max}$, it is possible to make the inverse transform $x^k=\Sigma_{s=1}^{k_{\max}}T_{ks}^{-1}\tilde{X}^s$ and also the change of variables, $\frac{\partial U}{\partial x^k}=\Sigma_{s=1}^{k_{\max}}T_{sk}\frac{\partial U}{\partial \tilde{X}^s}$, where the index s

$=1,2,\ldots,k_{\max}$ replaces the pair $\{\delta\alpha\}$. As a consequence, Eq. (46) is converted into the generalized Langevin equation for the coarse-grained degrees of freedom,

$$\ddot{\tilde{X}}^s = -\sum_{l=1}^{k_{\max}} V_{sl}\frac{\partial U}{\partial \tilde{X}^l} - \sum_{l=1}^{k_{\max}} \int \zeta_{sl}(t-\tau)\dot{\tilde{X}}^l(\tau)d\tau + \rho^s(t), \tag{47}$$

where

$$V_{sl} = \sum_{p,q=1}^{k_{\max}} T_{sp}\mu_{pq}^{-1}T_{lq}, \tag{48}$$

$$\zeta_{sl}(t) = \sum_{p,q=1}^{k_{\max}} T_{sp}\xi_{pq}(t)T_{ql}^{-1}, \tag{49}$$

$$\rho^s(t) = \sum_{l=1}^{k_{\max}} T_{sl}r^l(t). \tag{50}$$

Equation (47) provides a system of generalized Langevin equations for the coarse-grained degrees of freedom. The equations represent collective dynamics in a protein through a few interacting domains embedded in a dissipative medium. The equation can be entirely parametrized based on the dynamics of essential collective motions discussed in Sec. III. If the effective masses, mean forces, and memory kernels are available for a set of essential collective coordinates [see, e.g., Eqs. (29)–(31)], then the corresponding parameters $\partial U/\partial \tilde{X}^s$, V_{sl}, and $\zeta_{sl}(t)$ can be identified.

An important outcome of the theory is that the maximum number of addressable coarse-grained degrees of freedom, $\tilde{X}^s$, is equal to the number of essential collective coordinates, x^k. Thus, if $k_{\max}=3$, the corresponding set of three coarse-grained equations of motion can represent the $\{x,y,z\}$ coordinates for one domain or it can describe particular degrees of freedom belonging to two or three different domains. In a general case, the number of coarse-grained degrees of freedom cannot be larger than the number of essential collective coordinates $k_{\max}$ employed to identify the dynamic domains in the protein.

V. DISCUSSION

For more extensive applications of the present theory, all functions and parameters involved must be appropriately quantified. In particular, subject to such analysis are the following: (i) the effective mass matrix μ^{-1} in Eq. (29); (ii) the mean force associated with each of the essential collective coordinates, $-\partial U/\partial x^k$ in Eq. (30), or the potential of mean force U as a function of the collective coordinate x^k; (iii) the memory kernel matrix $\xi(t)$ in Eq. (31); (iv) the number of collective coordinates, $k_{\max}$, and the particular set of these coordinates.

The effective mass matrix μ^{-1} is the most straightforward to determine and is simply provided analytically by Eq. (29) for a predefined set of collective degrees of freedom. The mean force $-\partial U/\partial x^k$ and the memory kernel matrix $\xi_{kl}(t)$, in principle, can also be derived analytically employing Eqs. (20), (30), and (31). To accomplish this, however, the equation of motion (10) needs to be integrated to provide the force $\vec{F}$ as a function of time for the entire trajectory. Potentially, the theory developed in Sec. III can be employed as an advanced integrator for the molecular-dynamic trajectories given by Eq. (10), thereby providing a self-consistent solution of Eqs. (10) and (28). However, this requires a special formalism of integration, which still needs to be developed and tested. For practical purposes at the present stage of the theory, the mean force $-\partial U/\partial x^k$ and the memory matrix $\xi_{kl}(t)$ can be evaluated from molecular-dynamic trajectories. Thus, the potential of mean force is given by[9,19,21]

$$U(x^k) = -\beta^{-1}\ln[\Psi(x^k)], \tag{51}$$

where $\Psi(x^k)$ is the equilibrium phase-space density corresponding to the coordinate x^k. The requirement here is that the molecular-dynamic simulation must generate a canonical ensemble that satisfies the condition of ergodicity [9,19,21]. The memory kernel $\xi_{kl}(t)$ can also be derived from molecular-dynamics trajectories. The conventional way adopted in the literature employs various kinds of the memory equation for the velocity autocorrelation function $\langle \dot{x}^k(t)\dot{x}^k(0)\rangle$, which is derived from the generalized Langevin equation through a well-known procedure [15,21,28,59]. Two slightly different approaches have been reported. Thus, the procedure described in Ref. [59], when applied to Eq. (28), leads to the following set of equations for $\langle \dot{x}^k(t)\dot{x}^k(0)\rangle$,

$$\langle \dot{x}^k(t)\dot{x}^k(0)\rangle = -\sum_{l=1}^{k_{\max}} \mu_{kl}^{-1}\left\langle \frac{\partial U}{\partial x^l}x^k(0)\right\rangle - \sum_{l=1}^{k_{\max}} \int \xi_{kl}(t-\tau)\langle \dot{x}^l(\tau)x^k(0)\rangle d\tau, \tag{52}$$

whereas the approach adopted in Refs. [15,21] generates a different set of equations,

$$\langle \ddot{x}^k(t)\dot{x}^k(0)\rangle = -\sum_{l=1}^{k_{\max}} \mu_{kl}^{-1}\left\langle \frac{\partial U}{\partial x^l}\dot{x}^k(0)\right\rangle - \sum_{l=1}^{k_{\max}} \int \xi_{kl}(t-\tau)\langle \dot{x}^l(\tau)\dot{x}^k(0)\rangle d\tau. \tag{53}$$

The correlation functions $\langle \dot{x}^k(t)\dot{x}^k(0)\rangle$, $\langle \frac{\partial U}{\partial x^l}x^k(0)\rangle$, $\langle \frac{\partial U}{\partial x^l}\dot{x}^k(0)\rangle$, and $\langle \dot{x}^l(\tau)x^k(0)\rangle$ can be evaluated numerically from molecular-dynamic trajectories [60], and any of the sets of integral equations (52) and (53) can then be solved.

Identification of the number and particular set of the essential collective coordinates, $k_{\max}$, is another important point. In Sec. II it has already been noted that a physical criterion other than ranking of the mean-square displacements is required to identify the set of essential coordinates. A solution can be found if one recalls the major assumptions of this theory: (i) The displacements related to the essential motions are significantly larger than the fluctuations [to make the Taylor expansion in Eq. (11)], and (ii) the essential mo-

tions are significantly slower than the fluctuations [which makes Eq. (14) solvable to provide Eq. (15)].

Evidently, selection of the set of essential collective coordinates must be consistent with both assumptions. As has been shown in Sec. IV C, condition (i) is equivalent to requesting robustness of the domain system as a function of $k_{\max}$ combined with a standard ranking of the collective coordinates x^k according to the respective eigenvalues σ^k. However, this only determines the lowest possible value of $k_{\max}$, leaving unclear whether any limitations exist when $k_{\max}$ increases. A solution that emerges from the present theory is a complementary ranking according to the decay times τ_{xx} and τ_{ξ}, which correspond to the aucorrelation function $\langle x^k(t)x^k(0)\rangle$ and to the memory kernel, $\xi_{kk}(t)$, respectively, for each of the potentially essential coordinates x^k. Indeed, the criterion $\tau_{xx} > \tau_{\xi}$ in fact requires that motion along the corresponding collective coordinate is slow compared to fluctuations in the environment. This is consistent with results reported to date [20,21,28], according to which the decay time τ_{xx} is significantly larger than τ_{ξ} for major essential coordinates in proteins. Thus, the requirement $\tau_{xx} > \tau_{\xi}$ employed together with the standard ranking of the eigenvalues of the covariance matrix provides a sufficient criterion for identification of both lower and upper limits for the number of essential coordinates.

A related question is how long the total trajectory sampling time T should be to obtain consistent results. Thus, for any realistic sampling time T, there may exist slow modes that are not captured well because their characteristic time is larger than T [21]. These undersampled modes present a problem for the formalism developed here, since they combine small eigenvalues σ^k with large decay times τ_{xx} and thus do not fall either into the category of essential modes or into the category of fluctuations. A solution is to identify and eliminate the undersampled modes or example, through conventional drift reduction techniques.

Another aspect of trajectory sampling that is worth addressing is whether a trajectory that takes a time T should be described by a single set of collective coordinates or must there be a local subsampling over smaller time intervals $\Delta T_1, \Delta T_2, \ldots$ with respective subsets of essential coordinates. In the first case, the covariance matrix described by Eq. (1) needs to be computed for the entire trajectory time T, whereas in the second case the matrix has to be derived for the successive smaller time intervals $\Delta T_1, \Delta T_2, \ldots$ and the entire formalism needs to be applied to each interval individually. A related question is whether or not a set of collective coordinates identified for a given trajectory time T can be employed to describe dynamics of the protein over longer times. In the literature, pertinent results are scarce. According to Ref. [21], if a single collective coordinate is considered for the protein neurotensin, the average direction of this collective coordinate changes rather significantly with a characteristic time of the order of at least 10 ns. This indicates that the essential collective space changes with time along the phase trajectory, and thus caution is required in both identifying the appropriate time intervals for sampling as well as extrapolating the obtained results over longer times. In particular, a variability of the set of essential collective coordinates with time should result in dynamic changes in correlated domains that have been introduced in Sec. IV. The number of major domains, their size, composition, and interaction with the environment are expected to vary in the case if the set of essential collective coordinates changes.

The probable dynamic variability of correlated domains provides a straightforward approach to characterizing the stability of the protein's conformation, as well as presents a fundamental challenge for the theory of collective dynamics in proteins. The positive outcome is that the formalism of identification of domains introduced in Sec. IV offers a transparent and efficient methodology to verify whether the conformation space of a protein remains stable over a trajectory. If the number, size, content, and other dynamic characteristics of major domains do not depend significantly on the sampling interval, then the protein's conformation space can be considered to be largely stable. Otherwise, any significant changes in the conformation space would generate easily detectable changes in the domains. At the same time, the probable variability of domains presents a fundamental challenge for the formalism of protein dynamics, since the formalism needs to be developed further to include the variability of domains in a consistent way. Solving this major challenge appears to be one of most important and promising milestones in the future development of the theory of protein dynamics.

VI. SUMMARY

(i) This paper introduces a general formalism to derive the equations of motion for essential collective modes in a dynamic system that is assumed to be a protein molecule embedded in a solvent. Using the projection operator technique [11] a system of coupled generalized Langevin equations is derived for essential collective coordinates, which are generated by principal component analysis of molecular dynamic trajectories. The number of the essential degrees of freedom is not limited in the theory. Unlike other studies, the present theory is valid for any number of essential degrees of freedom. In particular, coupling of the degrees of freedom is described. The theory includes the model with a single essential degree of freedom, as a particular case.

(ii) An explicit analytic relation is established between the generalized Langevin equation of motion for the essential collective coordinates provided by PCA and that for the all-atom phase trajectory projected onto the subspace of essential collective degrees of freedom. Potentially, this relation allows the employment of the developed formalism as an advanced integrator for molecular dynamic trajectories.

(iii) The introduced formalism is applied to define correlated dynamic domains in a macromolecule. The domains are defined as groups of atoms that show a strong dynamic coupling in the generalized Langevin equation of motion and are identified through clustering of directional cosines of the essential collective degrees of freedom. Unlike other existing approaches to identify domains in proteins, no assumptions regarding the number of domains, their elementary building blocks, or interatomic interactions are made. No additional noise reduction is required, because the clustering is per-

formed in the space of essential collective motions where fluctuations are eliminated. Since the formalism of domain identification has been developed based on a rigorous theoretical background, the formalism is universal and physically transparent. An example of identification of dynamic domains is provided for protein G. The example demonstrates that the identified domains are composed of compact groups of atoms, although the spatial proximity of atoms is not required by the formalism. The identified domains show a reasonable match with the primary and secondary structure, but there is no complete similarity. Some domains contain entire elements of the secondary structure, others include only parts of such elements, and still others are composed of different elements that are located close to each other in the tertiary structure of the protein.

(iv) The role of the number of essential coordinates, $k_{\max}$, in defining correlated domains is analyzed. It is demonstrated that identification of domains with various essential dimensionalities $k_{\max}$ reflects complementary aspects of the domain structure. Using low dimensionalities reveals subtle features of the averaged motion in a particular projection, whereas higher dimensionalities, at which robustness of the domain system is achieved, provide a more complete and dynamically consistent description. For the example of protein G, robustness of the domain system with increasing essential dimensionality $k_{\max}$ is reached only with sufficiently high dimensionalities, $k_{\max} \sim 10$. This is consistent with the basic requirement of this theory, for the essential motion encompasses a major portion of the total displacement to provide a valid representation of protein's dynamics.

(v) It is demonstrated that the theory introduced has a strong potential to interpret experimental NMR measurements. The idea is to compare experimental NMR results with the outcomes of the theoretic clustering and defining the essential dimensionality at which the best match with experiment is reached. NMR-derived dynamic data that are only reproduced through low-dimensional clustering provide insight into the most delicate details of protein structure; however, only those features that are identified theoretically with sufficiently high essential dimensionalities would be fully representative of the coarse-grained dynamics in the macromolecule. An example of such an analysis is provided for protein G. It is shown that the symmetry of the domain system identified with the single essential coordinate, $k_{\max}=1$, resembles the long-range network of interstrand correlations extracted from the RDC analysis of backbone motions, whereas the robust domain system predicted with the set of ten collective coordinates, $k_{\max}=10$, matches the major regions of rigidity and softness in the conventional experimental profiles of the order parameter S^2. Thus, by employing different essential dimensionalities $k_{\max}$, two radically different sets of NMR measurements have been matched theoretically, the complementary nature of the measurements confirmed, and the physical meaning of the observed differences explained.

(vi) A distinguishing feature of the present identification of dynamic domains is that this approach is rigorously justified theoretically, which makes the results available for further theoretical developments. In this work, the dynamic correlated domains are employed as a starting point to construct an analytic coarse-grained model, which describes conformational motions in a protein through the system of interacting domains embedded in a dissipative medium. For the first time, generalized equations of motion for the Cartesian coordinates of the dynamic domains have been derived and parametrized analytically based on the equations of motion for the essential collective coordinates.

(vii) The physical requirements adopted in the formalism are discussed in detail. The major requirement is categorizing of collective modes either as essential degrees of freedom or as fluctuations. The essential displacements must be large and slow, whereas the fluctuations must be weak and high frequency. These requirements provide a physical criterion to identify essential degrees of freedom from the entire set of collective coordinates generated by PCA. In addition to the standard ranking of the mean-square displacements, the collective coordinates should be ranked according to the decay times of their autocorrelation functions. Another major requirement is that the number of essential collective coordinates cannot be less that the number of coarse-grained degrees of freedom that is desired to address.

(viii) Potential applications and further developments of the formalism are discussed. A fundamental challenge and one of most important and promising future milestones is the extension of the formalism to describe dynamic variability of correlated domains. In the present condition, the formalism offers a broad range of new opportunities to characterize and compare collective conformational behaviors in macromolecules based on their molecular dynamic trajectories, as well as for comparison of molecular-dynamic simulations with experiments and interpretation of experimentally derived dynamic information.

ACKNOWLEDGMENTS

The author thanks A. Kitao, A. Kobryn, A. Potapov, and A. Kovalenko for their helpful discussion of the work and M. Berjanskii for generating the molecular-dynamics trajectory for protein G. The images in Figs. 2, 4, and 7 were generated with VMD 1.8.3.

[1] M. M. Tatsuoka, *Multivariate Analysis* (Macmillian, New York, 1988).
[2] A. Kitao, F. Hirata, and N. Go, Chem. Phys. **158**, 447 (1991).
[3] A. E. Garcia, Phys. Rev. Lett. **68**, 2696 (1992).
[4] S. Hayward, A. Kitao, F. Hirata, and N. Go, J. Mol. Biol. **234**, 1207 (1993).
[5] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins **17**, 412 (1993).
[6] D. M. F. Van Aalten, B. L. De Groot, J. B. C. Findlay, H. J. C. Berendsen, and A. Amadei, J. Comput. Chem. **18**, 169 (1997).

[7] A. Kitao and N. Go, Curr. Opin. Struct. Biol. **9**, 174 (1999).
[8] See also comments in Sec. V.
[9] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).
[10] P. I. Diadone, M. D'Abramo, A. Di Nola, and A. Amadei, J. Am. Chem. Soc. **127**, 14825 (2005).
[11] See, for example, J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 3rd ed. (Academic Press, Amsterdam, 2006), Chap. 9, p. 255.
[12] Projection of the phase trajectory to obtain the essential motions associated with particular collective coordinates is implemented, for example, in the well-known code GROMACS, http://www.gromacs.org/.
[13] A. Amadei, B. L. de Groot, M.-A. Ceruso, M. Paci, A. Di Mola, and H. J. C. Berendsen, Proteins **35**, 283 (1999).
[14] P. Eastman, M. Pellegrini, and S. Doniach, J. Chem. Phys. **110**, 10141 (1999).
[15] D. E. Sagnella, J. E. Straub, and D. Thirumalai, J. Chem. Phys. **113**, 7702 (2000).
[16] A. Ansari, J. Chem. Phys. **112**, 2516 (2000).
[17] B. Oliva, X. Daura, E. Querol, F. X. Aviles, and O. Tapia, Theor. Chem. Acc. **105**, 101 (2000).
[18] L. Bu and J. E. Straub, Biophys. J. **85**, 1429 (2003).
[19] I. Kosztin, B. Barz, and L. Janosi, J. Chem. Phys. **124**, 064106 (2006).
[20] K. Moritsugu and J. C. Smith, J. Phys. Chem. **110**, 5807 (2006).
[21] O. F. Lange and H. Grubmüller, J. Chem. Phys. **124**, 214903 (2006).
[22] $Z(t)$ and $\vec{R}(t)$ in Eq. (15) are linear combinations of the terms $\sin(\omega_i t+\varphi_i)$, where ω_i^2 are the eigenvalues of $(1-P)\mathrm{m}^{-1}$ K.
[23] V. B. Magalinski, Sov. Phys. JETP **9**, 1381 (1959).
[24] E. Cortés, B. J. West, and K. Lindenberg, J. Chem. Phys. **82**, 2708 (1985).
[25] H.-X Zhou and R. Zwanzig, J. Phys. Chem. A **106**, 7562 (2002).
[26] G.-L. Ingold, *Lecture Notes in Physics*, edited by A. Buchleiter and K. Hornberher (Springer, Berlin, 2002), pp. 1–53.
[27] P. Hänggi and G.-L. Ingold, Chaos **15**, 026105 (2005).
[28] H. Kamberaj (2006, NINT-NRC, unpublished report).
[29] P. Hänggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys. **62**, 251 (1990).
[30] This is demonstrated, for example, by the fact that the effective masses μ that appear in Eq. (28) are representative of an averaged mass of atoms involved in the collective motion, rather than of a cumulative mass of any group of atoms. See also Eq. (34) with accompanying comments.
[31] S. O. Yesylevsky, V. N. Kharkyanen, and A. P. Demchenko, Biophys. J. **91**, 670 (2006).
[32] S. Hayward, A. Kitao, and H. Berendsen, Proteins **27**, 425 (1997).
[33] S. Hayward and H. Berendsen, Proteins **30**, 144 (1998).
[34] S. Hayward and R. A. Lee, J. Mol. Graphics Modell. **21**, 181 (2002).
[35] K. Hinsen, Proteins **33**, 417 (1998).
[36] K. Hinsen, A. Thomas, and M. J. Field, Proteins **34**, 369 (1999).
[37] A. Zhuravleva, D. M. Korzhnev, S. B. Nodle, L. E. Kay, A. S. Arseniev, M. Billeter, and V. Yu. Orekhov, J. Mol. Biol. **367**, 1079 (2007).
[38] The degenerated case of $E^k_{n,\alpha}=0$ for all k is not considered in this derivation; zero values for all direction cosines mean that motion of the domain is not captured by the set of the essential coordinates selected, and thus such a domain is not described by the respective equations of motion.
[39] The MD simulation, which has been kindly provided by M. Berjanskii, is discussed in more detail in our forthcoming publications.
[40] The selection of $k_{max}=10$ ensures a reasonable robustness of the domain system with a further increase of k_{max}, as discussed in more detail in Sec. IV C.
[41] A. K. Jain, M. N. Murty, and P. J. Flynn, ACM Comput. Surv. **3**, 264 (1999).
[42] Note that the interdomain distance d is a dimensionless value. This follows from the definition of the metric in the $3k_{max}$-dimensional space of directional cosines of essential collective degrees of freedom, where the domains are identified.
[43] In this paper, the distance d is normalized by $(3k_{max})^{1/2}$, in order to represent the variance of directional cosines per degree of freedom for various k_{max}.
[44] T. Konatsuzaki, K. Hoshino, Y. Matsunaga, G. J. Rylance, R. Johnston, and D. J. Wales, J. Chem. Phys. **122**, 084714 (2005).
[45] The intercluster distances that correspond to various $k_{\max}$ in Fig. 4, $d_1=0.0020$, $d_2=0.0026$, $d_5=0.0030$, $d_{10}=0.0039$, $d_{20}=0.0055$, and $d_{40}=0.0072$, were identified as the midpoint of the d value of the maximum curvature of the dependence, $N_{tot}(d)$, and that of the maximum of the function, $N_{tot}-N_1$ (see Fig. 1).
[46] A. G. Palmer, Annu. Rev. Biophys. Biomol. Struct. **30**, 129 (2001).
[47] D. A. Case, Acc. Chem. Res. **35**, 325 (2002).
[48] R. Brüschweiler, Curr. Opin. Struct. Biol. **13**, 175 (2003).
[49] A. Mittermaier and L. E. Kay, Science **312**, 224 (2006).
[50] V. A. Jarymowycz and M. J. Stone, Chem. Rev. (Washington, D.C.) **106**, 1624 (2006).
[51] M. L. Tillett, M. J. Blackledge, J. P. Derrick, L.-Y. Lian, and T. J. Norwood, Protein Sci. **9**, 1210 (2000).
[52] M. J. Stone, S. Gupta, N. Snyder, and Lynne Regan, J. Am. Chem. Soc. **123**, 185 (2001).
[53] D. Idiatullin, V. A. Daragan, and K. Mayo, J. Phys. Chem. **107**, 2602 (2003).
[54] J. B. Hall and D. Fushmann, J. Biomol. NMR **27**, 261 (2003).
[55] G. Bouvignies, P. Bernado, S. Meier, K. Cho, S. Grzesiek, R. Brüschweiler, and M. Blackledge, Proc. Natl. Acad. Sci. U.S.A. **102**, 13885 (2005).
[56] P. R. L. Markwick, G. Bouvignies, and M. Blackledge, J. Am. Chem. Soc. **129**, 4724 (2007).
[57] It can be seen that the residue numbering in Figs. 5(a) and 5(b) is somewhat different. In this work, the residues are numbered as in Fig. 5(a). To facilitate comparison with Fig. 5(b), most of the results discussed in Sec. IV are formulated in terms of the secondary structure rather than the numbers of residues.
[58] It should be noted that the analysis of rigidity and softness through the theoretical order parameter S_D is only applicable to dynamically consistent domain systems for sufficiently large k_{max} as described in Sec. IV C.
[59] M. Berkowitz, J. D. Morgan, D. J. Kouri, and J. Andrew McCammon, J. Chem. Phys. **75**, 2462 (1981).
[60] C. Xing and I. Andricioaei, J. Chem. Phys. **124**, 034110 (2006).