Condensed Matter Physics 2007, Vol. 10, No 3(51), pp. 441–457

# Towards coarse-grained modelling of proteins

M.Stepanova

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

Received May 31, 2007

This paper introduces a basic theoretical background to the description of conformational dynamics of proteins through a system of interacting domains. The essential collective degrees of freedom derived by principal component analysis of a molecular dynamics trajectory are used as dynamic variables defining the projection operator technique that underlies the formalism suggested. The explicit form of the corresponding projection operator is obtained, and the projection method is employed to derive systems of coupled generalized Langevin equations for both individual atomic degrees of freedom and essential collective degrees of freedom in a protein. A definition of correlated domains in proteins is introduced based on the analysis of the essential dynamics. Examples of identification of such domains are presented. A system of coupled generalized Langevin equations is derived representing the protein through a few interacting domains embedded into a dissipative medium. Further developments and potential applications of the formalism are outlined.

**Key words:** *protein dynamics, conformational changes, theory and modelling, projection operator, principal component analysis*

**PACS:** *87.15.He, 05.10.Gg, 87.15.Aa, 02.50.Sk*

## 1. Introduction

Building coarse-grained models for proteins is one of the major unresolved challenges for the theory. Proteins are complex soft-matter systems containing thousands of atoms interacting with each other within the protein molecule, as well as with the environment (solvent) and constantly changing their spatial conformations as a result of this interaction. It is known that the dynamics of protein molecules is highly hierarchical, e.g., it comprises chaotic high-frequency motions of individual atoms and small atomic groups superimposed with slower and more regular motions involving significant parts of the macromolecule. An appropriate identification, characterization, and prediction of these slow collective motions would tremendously benefit to both the basic understanding of protein dynamics and applications to the design of functional nanosystems employing proteins (drug design, biosensing, biodiagnostics, etc.).

A significant effort has been invested by researchers to develop a methodology of characterization of the collective modes in proteins through the multivariate analysis of molecular dynamic trajectories of proteins [1–6]. In this method, which is known as the principal component analysis (PCA), the trajectory of the protein in the phase space, $\vec{X}\,(t) = \{X_1\,(t)\,, X_2\,(t)\,, \dots X_{3N}\,(t)\}$, where $X_i$ are the Cartesian coordinates of individual atoms, is obtained from molecular dynamics simulations. The covariance matrix $c_{ij} = \langle (X_i\,(t) - \langle X_i\rangle)\ \ (X_j\,(t) - \langle X_j\rangle)\rangle$ is defined through averaging over the entire trajectory, and the eigenvalue problem,

$$\sum_{j=1}^{3N} c_{ij} E_j^k = \sigma^k E_i^k, \qquad i = 1, 2, \dots 3N \tag{1}$$

is solved to obtain the eigenvectors $\vec{E}^k = \left\{E_1^k, E_2^k, \dots E_{3N}^k\right\}$ and eigenvalues $\sigma^k$ $(k = 1, 2, \dots 3N)$, where $N$ is the total number of atoms considered. The normalized eigenvectors, $\vec{E}^1, \vec{E}^2, \dots \vec{E}^{3N}$, represent a set of $3N$ orthogonal collective degrees of freedom. One can consider the set of eigenvectors as an intrinsic coordinate frame in the phase space, and project the phase trajectory $\vec{X}\,(t)$

on them,

$$\left(\vec{X}\left(t\right)\cdot\vec{E}^{k}\right)=\sum_{i=1}^{3N}E_{i}^{k}X_{i}\left(t\right)=x^{k}\left(t\right),\qquad k=1,2,\ldots 3N. \tag{2}$$

The functions $x^k(t)$ defined by equation (2) can be viewed as the collective coordinates representing the conformational changes in the protein embedded in solvent. The eigenvalues $\sigma^k$ derived from equation (1) represent the mean square displacements along the corresponding collective degrees of freedom. Thus, it is possible to rank the collective degrees of freedom according to the magnitude of the associated eigenvalues, and to consider a truncated set of collective coordinates, $x^1, x^2, \ldots x^{k_{\max}}$, $k_{\max} \ll 3N$, which include only those collective coordinates that have the highest magnitude of the displacements [1–7]. This truncated set of collective coordinates is sometimes referred to as the essential degrees of freedom [4,7]. The complementary set of collective coordinates, $x^{k_{\max}+1}, x^{k_{\max}+2}, \ldots x^{3N}$ is interpreted as small-amplitude fluctuations.

The PCA-based techniques have provided valuable information about the geometry of the conformational changes, and are currently implemented in popular simulation packages such as AMBER and GROMACS. However, this formalism alone does not provide a sufficient insight into the dynamics of the conformational motions in macromolecules. Thus, 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 analysing the geometry of the phase trajectory alone. Thus, dynamics of the collective motions needs to be addressed in addition to their characterization through the standard PCA techniques.

Extensive attempts to develop a dynamic theory of conformational motions in proteins are presented in references [3,7–16] and citations therein. It has been suggested to describe the dynamics of proteins by the classic Langevin equations of motion, employing either the essential collective coordinates $x$ [3,14] or the Cartesian coordinates of atoms $X$ [9] as dynamic variables. Several authors proposed employing the generalized Langevin equation as a more comprehensive approach to the description of damped motion of individual atoms in a protein [10,12,13]. In the recent study [16], it has been postulated that the motion along the essential 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 of a molecular dynamics trajectory [16]. However, applicability of the generalized Langevin equation to the essential degrees of freedom extracted from molecular-dynamic trajectories has not been proven rigorously, and any relation between the Langevin equation for Cartesian coordinates of atoms in the protein and those for the collective essential coordinates derived through PCA has not been established. Furthermore, employing a single collective degree of freedom is too a restrictive assumption for realistic proteins.

A fundamental unsolved problem that requires accounting for more than one collective degree of freedom in proteins, is representation of the collective motion in terms of particular domains containing atoms that move in a coherent way. Efforts in identifying such domains based on molecular simulations have been recently reviewed [17]. Difficulties arise even with the very definition of domains, which sometimes include rather vague criteria such as being a visually recognizable substructure in the protein [17]. In the most elaborate approach [18–21], 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 those elementary building blocks should be postulated a *priori* (residuals, groups of a few atoms, etc.). Also the differences in motion that need to be captured are very subtle and susceptible to thermal noise, to sampling scheme, and to other uncertainties. A proper filtering of these unwanted impacts is complicated and computationally expensive to implement. A universal and dynamically justified concept for identifying the correlated domains has not been developed to the date. This challenge might be solved through a theoretical approach employing collective coordinates as dynamic variables in a protein, and identifying the couplings that cause the formation of correlated domains. However, a theoretical methodology describing the conformational dynamics in a protein based on multiple collective degrees of freedom still needs to be developed to this end.

In this paper, a comprehensive dynamic formalism is introduced that eventually permits to define the correlated domains in a protein, and describe their motion by a system of dynamic equations of motion that are parameterized employing PCA of molecular dynamics trajectories. In section 1, the set of essential collective coordinates derived by PCA is employed to construct the projection operator. In section 2, systems of coupled generalized equations for both individual atomic degrees of freedom and essential collective degrees of freedom are derived through the PCA-based projection operator method. In section 3, a definition of correlated domains in proteins is introduced based on the analysis of coupling of dynamic variables in equations of motion, and the examples of identifying such domains are presented. In section 4, generalized Langevin equations are derived describing the protein as a system of interacting domains. Potential applications of the formalism, further developments, and related challenges are discussed. Section 5 summarizes conclusions of the work.

## 2. Projection operator methodology for protein dynamics

In this section, systems of coupled generalized Langevin equations are derived for both Cartesian degrees of freedom of individual atoms and the collective essential coordinates in a protein, based on the projection operator method [22–25]. For this purpose, the projection operator is defined employing a set of multiple essential degrees of freedom, which are derived through the PCA of molecular dynamics trajectories.

### 2.1. The projection operator from PCA-defined eigenvectors

Consider the eigenvectors $\vec{E}^1, \vec{E}^2, \ldots \vec{E}^{3N}$ derived from equation (1). As it has been discussed, the eigenvectors represent the set of the protein's collective degrees of freedom, which can be subdivided into the essential degrees of freedom describing significant displacements, $\vec{E}^1, \vec{E}^2, \ldots \vec{E}^{k_{\max}}$, $k_{\max} \ll 3N$, and the complementary set of collective degrees of freedom that correspond to small-amplitude fluctuations, $\vec{E}^{k_{\max}+1}, \vec{E}^{k_{\max}+2}, \ldots \vec{E}^{3N}$. The number of essential degrees of freedom, $k_{\max}$, is not determined or limited at this point. The set of essential degrees of freedom and that associated with fluctuations can be viewed as two orthogonal subspaces of the $3N$-dimensional phase space.

Next, let us introduce the operator $P$ and the complementary operator $1-P$,

$$P\vec{Y} = \sum_{k=1}^{k_{\max}} \left(\vec{E}^k \cdot \vec{Y}\right) \vec{E}^k, \qquad (1-P)\,\vec{Y} = \sum_{k=k_{\max}+1}^{3N} \left(\vec{E}^k \cdot \vec{Y}\right) \vec{E}^k, \tag{3}$$

where $\vec{Y}$ is an arbitrary vector. One can easily check that the operators $P$ and $1-P$ can be interpreted as geometrical projections of the vector $\vec{Y}$ onto the subspace of the essential degrees of freedom and onto the subspace of fluctuations, respectively. This can be demonstrated by the following examples,

$$\begin{aligned} &PP\vec{Y} = P\vec{Y}, && P\,(1-P)\,\vec{Y} = 0, \\ &(1-P)\,P\vec{Y} = 0, && (1-P)\,(1-P)\,\vec{Y} = (1-P)\,\vec{Y}. \end{aligned} \tag{4}$$

In particular, when the operators $P$ and $1-P$ are applied to the trajectory vector $\vec{X}(t) = \{X_1(t), X_2(t), \ldots X_{3N}(t)\}$, this provides the essential component of the trajectory $\vec{X}^{\mathrm{E}}$ and the fluctuation component $\vec{X}^{1-\mathrm{E}}$,

$$\begin{aligned} P\vec{X} &= \vec{X}^{\mathrm{E}}, \\ (1-P)\,\vec{X} &= \vec{X}^{1-\mathrm{E}}. \end{aligned} \tag{5}$$

It is clear that

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

### 2.2. Essential dynamics for individual coordinates of atoms

The next step is employing the introduced projection operators to derive the generalized Langevin equations for individual atomic degrees of freedom, using the analogy with the Mori projection operator formalism [22–25]. Consider again the trajectory of the entire system, $\vec{X}\left(t\right) = \left\{X_1\left(t\right), X_2\left(t\right), \dots X_{3N}\left(t\right)\right\}$. The vector $\vec{X}\left(t\right)$ obeys the equation of motion,

$$\ddot{\vec{X}} = m^{-1}\vec{F}\left(\vec{X}\right), \tag{7}$$

where $m$ is a diagonal matrix providing the masses of atoms. Taking into account equation (6) where $\vec{X}^{1-\mathrm{E}}$ represents minor changes in the positions of atoms as compared to a more pronounced essential motion given by $\vec{X}^{\mathrm{E}}$, it is possible to approximately represent the force in the right-hand side of equation (7) in the following form,

$$\vec{F}\left(\vec{X}\right) = \vec{F}\left(\vec{X}^{\mathrm{E}} + \vec{X}^{1-\mathrm{E}}\right) \approx \vec{F}\left(\vec{X}^{\mathrm{E}}\right) + \frac{\partial \vec{F}\left(\vec{X}^{\mathrm{E}}\right)}{\partial \vec{X}^{\mathrm{E}}}\vec{X}^{1-\mathrm{E}} = \vec{F}\left(\vec{X}^{\mathrm{E}}\right) - K\vec{X}^{1-\mathrm{E}}. \tag{8}$$

In equation (8), $\vec{F}\left(\vec{X}^{\mathrm{E}}\right)$ is the mean force, $K\vec{X}^{1-\mathrm{E}}$ represents fluctuations of the force, and $K$ is the matrix with the elements $K_{ij} = -\partial F_i / \partial X_j$. Equation (7) can thus be replaced with,

$$\ddot{\vec{X}} = m^{-1}\left(\vec{F}\left(\vec{X}^{\mathrm{E}}\right) - K\vec{X}^{1-\mathrm{E}}\right). \tag{9}$$

Next, the projection operators $P$ and $1-P$ as defined in section 2.1 are applied to both sides of equation (9), which gives the formal equations of motion for $\vec{X}^{\mathrm{E}}$ and $\vec{X}^{1-\mathrm{E}}$, respectively:

$$\ddot{\vec{X}}^{\mathrm{E}} = Pm^{-1}\left(\vec{F}\left(\vec{X}^{\mathrm{E}}\right) - K\vec{X}^{1-\mathrm{E}}\right), \tag{10}$$

$$\ddot{\vec{X}}^{1-\mathrm{E}} = \left(1-P\right)m^{-1}\left(\vec{F}\left(\vec{X}^{\mathrm{E}}\right) - K\vec{X}^{1-\mathrm{E}}\right). \tag{11}$$

If the fluctuation contribution, $\vec{X}^{1-\mathrm{E}}$, changes much faster than $\vec{X}^{\mathrm{E}}$, the elements of the matrix $K$ can be considered as constants. With this assumption, equation (11) can be solved with respect to $\vec{X}^{1-\mathrm{E}}$, which gives the general solution in the form,

$$\vec{X}^{1-\mathrm{E}}\left(t\right) = \int\limits_o^t Z\left(t-\tau\right)\left(1-P\right)m^{-1}\vec{F}\left(\vec{X}^{\mathrm{E}}\left(\tau\right)\right)\mathrm{d}\tau + \vec{R}\left(t, \vec{X}^{1-\mathrm{E}}\left(0\right), \dot{\vec{X}}^{1-\mathrm{E}}\left(0\right)\right). \tag{12}$$

The kernel $Z\left(t\right)$ under the integral in the right-hand side of equation (12) in a general case is a non-diagonal matrix, whose elements are linear combinations of the terms $\sin\left(\omega_i t + \varphi_i\right)$, where $\omega_i^2$ are the eigenvalues of $\left(1-P\right)m^{-1}K$. The second term in the right-hand side of equation (12), which is symbolically represented by $\vec{R}$, is also a linear combination of the harmonic functions of time such as $\sin\left(\omega_i t + \varphi_i\right)$, weighted with the values of $\vec{X}^{1-\mathrm{E}}\left(0\right)$ and $\dot{\vec{X}}^{1-\mathrm{E}}\left(0\right)$ at the initial time $t = 0$. $\vec{R}$ can be viewed as the contribution of random noise to $\vec{X}^{1-\mathrm{E}}$.

The expression $\left(1-P\right)m^{-1}\vec{F}\left(\vec{X}^{\mathrm{E}}\right)$ under the integral in the right-hand side of equation (12) represents the projection of the mass-weighted force $m^{-1}\vec{F}\left(\vec{X}^{\mathrm{E}}\right)$ acting in the subspace of essential motions, onto the subspace of fluctuations. This can be rephrased as the coupling of fluctuations with the essential degrees of freedom. In the case of bilinear coupling, the solution of equation (12) can be represented in the following form [26–31],

$$\vec{X}^{1-\mathrm{E}}\left(t\right) = Z_{\mathrm{H}}\left(0\right)\vec{X}^{\mathrm{E}}\left(t\right) - Z_{\mathrm{H}}\left(t\right)\vec{X}^{\mathrm{E}}\left(0\right) - \int\limits_o^t Z_{\mathrm{H}}\left(t-\tau\right)\dot{\vec{X}}^{\mathrm{E}}\left(\tau\right)\mathrm{d}\tau + \vec{R}_{\mathrm{H}}\left(t\right). \tag{13}$$

The form of the damping kernel $Z_{\mathrm{H}}(t)$ and the random function $\vec{R}_{\mathrm{H}}(t)$ for the case of bilinear coupling can be found, for example, in references [27] and [30]. Substitution of equation (13) in the right-hand side of equation (10) provides the equation of motion for the projection of the trajectory onto the subspace of essential motions $\vec{X}^{\mathrm{E}}(t)$,

$$\begin{aligned}\ddot{\vec{X}}^{\mathrm{E}} \;=\;& Pm^{-1}\left[\vec{F}\left(\vec{X}^{\mathrm{E}}\right)+KZ_{\mathrm{H}}(0)\,\vec{X}^{\mathrm{E}}(t)-K\int\limits_{o}^{t}Z_{\mathrm{H}}(t-\tau)\dot{\vec{X}}^{\mathrm{E}}(\tau)\,\mathrm{d}\tau\right.\\ &\left.-KZ_{\mathrm{H}}(t)\,\vec{X}^{\mathrm{E}}(0)+K\vec{R}_{\mathrm{H}}(t)\right], \end{aligned}\tag{14}$$

which can be converted into the form similar to the generalized Langevin equation,

$$\ddot{\vec{X}}^{\mathrm{E}}=Pm^{-1}\left[-\frac{\partial U\left(\vec{X}^{\mathrm{E}}\right)}{\partial\vec{X}^{\mathrm{E}}}-\int\limits_{o}^{t}Z_{X}(t-\tau)\dot{\vec{X}}^{\mathrm{E}}(\tau)\,\mathrm{d}\tau+\vec{R}_{X}(t)\right].\tag{15}$$

Here $U\left(\vec{X}^{\mathrm{E}}\right)$ is the potential of mean force, defined in such a way that $\partial U\left(\vec{X}^{\mathrm{E}}\right)/\partial\vec{X}^{\mathrm{E}} = -\vec{F}\left(\vec{X}^{\mathrm{E}}\right)-KZ_{\mathrm{H}}(0)\,\vec{X}^{\mathrm{E}}(t)$; the integral $-\int_{o}^{t}Z_{X}(t-\tau)\dot{\vec{X}}^{\mathrm{E}}(\tau)\,\mathrm{d}\tau$ represents the dissipative force with the memory kernel $Z_{X}(t)=KZ_{\mathrm{H}}(t)$; and $\vec{R}_{X}(t)=-KZ_{\mathrm{H}}(t)\,\vec{X}^{\mathrm{E}}(0)+K\vec{R}_{\mathrm{H}}(t)$ can be interpreted as an external random force, in the sense that $\vec{R}_{X}(t)$ does not depend on the dynamics of the system considered [25,26] and satisfies the requirements $\langle R_{X,i}(t)\rangle=0$ and $\langle R_{X,i}(0)\,R_{X,j}(t)\rangle = \beta^{-1}Z_{X,ij}(t)$ [12,30].

The final step is rewriting equation (15) in the form of the set of scalar equations of motion for $3N$ atomic coordinates $X_i^{\mathrm{E}}$:

$$\ddot{X}_i^{\mathrm{E}}(t)=-\sum_{j=1}^{3N}C_{ij}m_j^{-1}\frac{\partial U}{\partial X_j^{\mathrm{E}}}-\sum_{l=1}^{3N}\int\limits_{o}^{t}\Xi_{il}(t-\tau)\dot{X}_l^{\mathrm{E}}(\tau)\,\mathrm{d}\tau+\rho_i(t)\,,\qquad i=1,2,\ldots 3N.\tag{16}$$

Here

$$C_{ij}=\sum_{k=1}^{k_{\max}}E_i^kE_j^k;\tag{17}$$

and

$$\Xi_{il}(t)\;=\;\sum_{j=1}^{3N}C_{ij}m_j^{-1}Z_{X,lj}(t),\qquad \rho_i(t)=\sum_{j=1}^{3N}C_{ij}m_j^{-1}R_{X,j}(t)\,.\tag{18}$$

The system of equations of motion (16) describes the trajectories for all atoms in the system, projected onto the subspace of the essential degrees of freedom that are identified by PCA of atomic trajectories. The first term in the right-hand side represents a “purely” essential motion defined by the mean forces $-\partial U/\partial X_j^{\mathrm{E}}$. The other terms in the right-hand side describe the effect of fluctuations onto the essential motion. The fluctuations are included in the equation in the form of dissipative force $-\int_{o}^{t}\Xi_{il}(t-\tau)\dot{X}_l^{\mathrm{E}}(\tau)\,\mathrm{d}\tau$ and the random force $\rho_i(t)$.

It can be seen that in equation (16), the atomic degrees of freedom are coupled through the summations in the right-hand side. Intuitively, the coupling of the degrees of freedom of individual atoms should be responsible for the formation of coherent domains in proteins and therefore, this coupling is of particular interest in the context of this paper. According to equation (16), the coupling between the particular degrees of freedom $i$ and $j$ is defined, first of all, by the coefficients

$C_{ij}$. Following equation (17), the coefficients $C_{ij}$ are determined by the values $E_i^k$, which represent the direction cosines of the eigenvectors $\vec{E}^k$ in the $3N$-dimentional phase space of the protein, $\vec{E}^k = \left\{E_1^k, E_2^k, \ldots E_{3N}^k\right\}$. The values $E_i^k$ can also be viewed as the projections of the eigenvectors $\vec{E}^k$ onto the Cartesian degrees of freedom of individual atoms. It is noteworthy that only the eigenvectors that correspond to the essential degrees of freedom ($k \leqslant k_{\max}$) contribute to the coupling. Clearly, the essential collective coordinates should play a central role in building coarse-grained models of proteins. To better understand this role, equations of motions for the collective coordinates $x^k$ will now be derived and analysed.

### 2.3. Essential dynamics for collective coordinates and coupling of atomic degrees of freedom

By the definition, the essential collective coordinates $x^k$ can be obtained through the scalar product $\left(\vec{X}^{\mathrm{E}} \cdot \vec{E}^k\right) = x^k$. Accordingly, the equations of motion for $x^k\left(t\right)$ are provided by a similar procedure applied to both sides of equation (15),

$$\left(\ddot{\vec{X}}^{\mathrm{E}} \cdot \vec{E}^k\right) = \ddot{x}^k = \left(\vec{E}^k \cdot Pm^{-1}\left[-\frac{\partial U\left(\vec{X}^{\mathrm{E}}\right)}{\partial \vec{X}^{\mathrm{E}}} - \int\limits_o^t Z_X\left(t-\tau\right)\dot{\vec{X}}^{\mathrm{E}}\left(\tau\right)\mathrm{d}\tau + \vec{R}_X\left(t\right)\right]\right). \tag{19}$$

To simplify this equation, let us first note that for an arbitrary vector $\vec{Y}$,

$$\left(\vec{E}^k \cdot P\vec{Y}\right) = \sum_{l=1}^{k_{\max}}\left(\vec{E}^k \cdot \vec{E}^l\right)\left(\vec{E}^l \cdot \vec{Y}\right) = \left(\vec{E}^k \cdot \vec{Y}\right),$$

and, therefore, the operator $P$ in the right-hand side of equation (19) can be omitted. Second, since $\vec{X}^{\mathrm{E}} = \sum\limits_{k=1}^{k_{\max}} \vec{E}^k x^k$, the change of variables $\partial U/\partial \vec{X}^{\mathrm{E}} = \sum\limits_{k=1}^{k_{\max}} \vec{E}^k \partial U/\partial x^k$ can be employed. With these improvements, equation (19) can be rewritten as follows,

$$\ddot{x}^k = -\sum_{l=1}^{k_{\max}} \mu_{kl}^{-1}\frac{\partial U}{\partial x^l} - \sum_{l=1}^{k_{\max}} \int\limits_0^t \xi_{kl}\left(t-\tau\right)\dot{x}^l\left(\tau\right)\mathrm{d}\tau + r^k\left(t\right), \qquad k = 1, 2, \ldots k_{\max}; \tag{20}$$

where

$$\mu_{kl}^{-1} = \sum_{i=1}^{3N} E_i^k E_i^l m_i^{-1}, \tag{21}$$

$$\xi_{kl}\left(t\right) = \sum_{i,j=1}^{3N} E_i^k E_j^l m_i^{-1} Z_{X,ij}\left(t\right), \tag{22}$$

$$r^k\left(t\right) = \sum_{i=1}^{3N} E_i^k m_i^{-1} R_{X,i}\left(t\right). \tag{23}$$

Equation (20) provides a system of equations of motion for the essential collective coordinates $x^k$. It is evident that in a general case, any essential collective coordinate $x^k$ is dynamically coupled to other essential collective coordinates in the system. The coupling of the mean forces is provided by the matrix of effective reciprocal mass $\mu^{-1}$, whose elements are given by equation (21), and the coupling of the dissipative forces is provided by the memory matrix $\xi\left(t\right)$, as given by equation (22).

The meaning and structure of elements of the matrices $\mu^{-1}$ and $\xi\left(t\right)$ are easier to understand, considering the simplified case of a single essential degree of freedom, when $k_{\max}=1$. In this case, the effective reciprocal mass is simply

$$\mu^{-1} = \sum_{i=1}^{3N}\left(E_i^1\right)^2 m_i^{-1}, \tag{24}$$

which can be interpreted as a weighted average of the reciprocal masses of atoms, with the weights equal to $\left(E_i^1\right)^2$. Thus, the value $\left(E_i^1\right)^2$ represents a measure of involvement of the $i$-th atomic degree of freedom in the collective dynamics of the system. The memory function is given by a weighted summation of the contributions from individual atomic degrees of freedom as well,

$$\xi\left(t\right) = \sum_{i,j=1}^{3N} E_i^1 E_j^1 m_i^{-1} Z_{X,ij}\left(t\right) = \sum_{i=1}^{3N} \left(E_i^1\right)^2 m_i^{-1} Z_{X,ii}\left(t\right) + \sum_{i\neq j} E_i^1 E_j^1 m_i^{-1} Z_{X,ij}\left(t\right). \tag{25}$$

In the right-hand side of equation (25), the most significant term (containing the diagonal elements $Z_{X,ii}$) has the structure similar to that in equation (24), e.g. involvement of $i$-th atomic degree of freedom in the collective dynamics is characterized by the value $\left(E_i^1\right)^2$.

In a general case of multiple essential collective degrees of freedom, the values of the directional cosines $E_i^k$ can be interpreted as a measure of correlation of a particular atomic degree of freedom $i$ with the essential collective coordinate $k$. Clearly, the direction cosines $E_i^k$ can adopt positive, negative, or zero values. In the first case, the collective mode represented by $\vec{E}^k$ and the atomic degree of freedom $i$ are in phase, in the second case they are in anti-phase, and in the third case there is no correlation. The magnitude $\left|E_i^k\right|$ is representative of the level of the correlation; the larger $\left|E_i^k\right|$ is, the stronger is the involvement of the atomic degree of freedom $i$ into the collective mode $k$.

One can conclude that the coupling of individual atomic degrees of freedom that is discussed in section 2.2 is mediated by correlations of the atomic degrees of freedom with the essential collective degrees of freedom in the system. This fact is reflected by the structure of the coupling coefficients $C_{ij}$ which, after equation (17), are given by a sum of the expressions $E_i^k E_j^k$ over all essential collective modes $k$.

## 3. Correlated domains in macromolecules

Having introduced the methodology to the description of essential dynamics in macromolecules, the next step is attempting to represent this dynamics through domains containing the atoms that move in a coherent way, e.g. show a strong coupling. In section 2 it has been demonstrated that essential collective coordinates obtained through PCA of molecular dynamics trajectories play an important role as mediators of coupling of individual atomic degrees of freedom. However, despite a rather common expectation, the essential collective coordinates generally do not explicitly represent any particular groups of atoms. This is demonstrated, for example, by the fact that the elements of the matrix of reciprocal effective mass $\mu_{ij}^{-1}$ in equation (20) are representative of a weighted average of masses of atoms involved in the collective motion, rather than of a cumulative mass of any group of atoms. Below, coherently moving domains of atoms in a protein are identified based on the analysis of couplings of atomic degrees of freedom. This approach does not employ any *a priori* assumptions regarding the structure of domains, and is based on the analysis of the coupling coefficients $C_{ij}$ in equation (16).

### 3.1. Correlated domains from dynamic coupling of coordinates of atoms

Consider the set of equations of motion for projected atomic coordinates (16), which we rewrite as follows,

$$\ddot{X}_i^{\mathrm{E}} = \sum_{j=1}^{3N} C_{ij} Y_j\,, \tag{26}$$

where $i = 1, 2, \ldots 3N$, and

$$Y_j = -m_j^{-1}\frac{\partial U}{\partial X_j^{\mathrm{E}}} - m_j^{-1}\sum_{l=1}^{3N}\int_o^t Z_{X,lj}\left(t-\tau\right)\dot{X}_l^{\mathrm{E}}\left(\tau\right)\mathrm{d}\tau\, + m_j^{-1}R_{X,j}\left(t\right).$$

Consider again the essential eigenvectors $\vec{E}^k = \left\{E_1^k, E_2^k, \ldots E_{3N}^k\right\}$, whose direction cosines $E_i^k$ define the coupling coefficients $C_{ij}$ as given by equation (17). In a system containing $N$ atoms, the entire sets of direction cosines $\left\{E_1^k, E_2^k, \ldots E_{3N}^k\right\}$ can be subdivided into $N$ subsets each containing 3 values $\left\{E_{n,x}^k, E_{n,y}^k, E_{n,z}^k\right\}$, where n = 1,... N. Each of these subsets represents the direction cosines relative to the Cartesian degrees of freedom of an individual atom $x$, $y$, and $z$. The coupling coefficients $C_{ij}$ can now be represented by

$$C_{n_1,\alpha,n_2,\beta} = \sum_{k=1}^{k_{\max}} E_{n_1,\alpha}^k E_{n_2,\beta}^k\,, \tag{27}$$

where $n_1,\ n_2 = 1, 2, \ldots N$; $\alpha,\ \beta = 1, 2,$ or 3 denote the Cartesian degrees of freedom $x$, $y$, and $z$; and $E_{n,\alpha}^k$ are the directional cosines of the collective vectors $\vec{E}^k$ with respect to the atomic degrees of freedom $\{n, \alpha\}$. Accordingly, equation (26) is converted to

$$\ddot{X}_{n_1,\alpha}^{\mathrm{E}} = \sum_{n_2=1}^{N}\sum_{\beta=1}^{3} C_{n_1,\alpha,n_2,\beta}\, Y_{n_2,\beta} = \sum_{n_2=1}^{N}\sum_{\beta=1}^{3}\left(\sum_{k=1}^{k_{\max}} E_{n_1,\alpha}^k E_{n_2,\beta}^k\right) Y_{n_2,\beta}\,. \tag{28}$$

As it has been previously discussed, the values $E_{n,\alpha}^k$ represent correlations of the essential collective degrees of freedom with individual atomic degrees of freedom, and at the same time, they define the couplings between the degrees of freedom of individual atoms. Therefore, 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 $k$, and are nonzero for at least some $k$ [32]. This definition of domains can be illustrated by the following simple classification of cross-correlation terms in the equation of motion for the atomic coordinates in a hypothetic protein containing a single correlated domain:

(i) The atoms $n_1$ and $n_2$ belong to the correlated domain and their Cartesian degrees of freedom are similar ($\alpha = \beta$). Since in this case $E_{n_1,\alpha}^k \approx E_{n_2,\beta}^k$ for all $k$, and $E_{n_1,\alpha}^k \neq 0$ for at least some $k$, the coupling coefficients in the equation of motion for the atomic coordinate $\{n_1, \alpha\}$ are nonzero and positive. Let us denote such cases by

$$\left.\ddot{X}_{n_1,\alpha}^{\mathrm{E}}\right|_{(i)} = \sum_{n_2 \in \{N^\delta\}} C_{n_1,\alpha,n_2,\alpha} Y_{n_2,\alpha}. \tag{29}$$

In equation (29) $\delta$ denotes the domain, $\left\{N^\delta\right\}$ denotes the set of atoms in the domain $\delta$, and the expression $n_2 \in \left\{N^\delta\right\}$ means that the atom $n_2$ belongs to the domain $\delta$.

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

$$\left.\ddot{X}_{n_1,\alpha}^{\mathrm{E}}\right|_{(ii)} = \sum_{n_2=1}^{N}\sum_{\beta\neq\alpha} C_{n_1,\alpha,n_2,\beta} Y_{n_2,\beta} + \sum_{n_2 \notin \{N^\delta\}} C_{n_1,\alpha,n_2,\alpha} Y_{n_2,\alpha}. \tag{30}$$

Since the values $E_{n_1,\alpha}^k$ and $E_{n_2,\beta}^k$ can now vary in both magnitude and sign, the coupling coefficients $C_{n_1,\alpha,n_2,\beta}$ and $C_{n_1,\alpha,n_2,\alpha}$ in equation (30) are given by a summation of both positive and negative terms. This generates significantly smaller values for the coupling coefficients, which can also vary in signs. As a result, the total contribution to the equation of motion for the coordinate $\{n_1, \alpha\}$ is much less than in the case (i).

The entire equation of motion for the atomic coordinate $\{n_1, \alpha\}$ in a single correlated domain reads,

$$\ddot{X}_{n_1,\alpha}^{\mathrm{E}} = \left.\ddot{X}_{n_1,\alpha}^{\mathrm{E}}\right|_{(i)} + \left.\ddot{X}_{n_1,\alpha}^{\mathrm{E}}\right|_{(ii)}, \tag{31}$$

where the first term in the right-hand side describes the coupling of degrees of freedom which are correlated within the domain, whereas the second term corresponds to the coupling of degrees of freedom which do not show a strong correlation within the domain. Since in most cases the contribution (i) is much larger than (ii), in the first approximation one can disregard the uncorrelated term, and assume that $\ddot{X}^{\mathrm{E}}_{n_1,\alpha} \approx \ddot{X}^{\mathrm{E}}_{n_1,\alpha}\Big|_{(i)}$.

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 equation of motion for projected atomic coordinates. The domains are identified as groups of atoms, for which the corresponding direction cosines of the essential collective degrees of freedom $E^k_{n,\alpha}$ adopt similar values for each $k$. No assumption regarding any elementary building blocks and/or interatomic interactions are made in this work to identify the domains.

### 3.2. Example: domains in a misfolded prion protein

In this section, correlated domains in a macromolecule are identified for the example of a misfolded conformation of prion protein PrP 27–30, which has been recently suggested as a heuristic prototype of the pathogenic prion isoform apt to aggregate into oligomers [33]. The structure of a monomer of PrP 27–30, which contains 139 residues, is shown in Fig. 1. The molecule contains two $\alpha$-helices, as well as several layers of $\beta$-strands. To the date, little is known about the properties of such misfolded prion isoforms, as well as about the mechanism through which they arise and aggregate into dimmers, trimmers, and other oligomers. Thus, a rigorous and efficient approach to the characterization of structure and stability of the misfolded prion proteins would be of a tremendous importance.

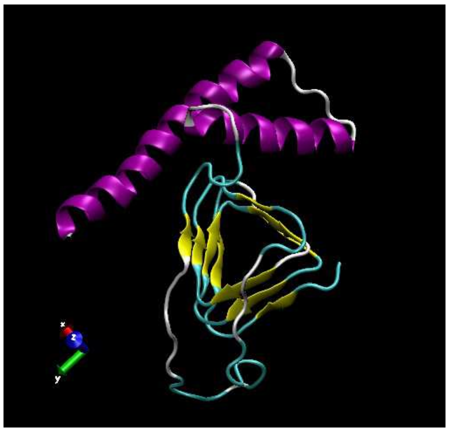


**Figure 1.** The secondary structure of a monomer of PrP 27–30 [33].

Starting with the structure shown in figure 1, a molecular dynamics trajectory of the solvated protein was generated using the NAMD2 code with Amber parm99 force field and explicit TIP3P water [34]. After equilibration, a trajectory of more than 3 ns in duration has been obtained. At the beginning of the trajectory as well as after every 1 ns, 0.2 ns intervals were analysed by PCA. Four such intervals have been analysed independently. For essential collective coordinates, 30 principal components with the highest eigenvalues ($k_{\max} = 30$) were identified for each of these four intervals. All atoms in the molecule were accounted for when doing the PCA, which corresponded to $N = 2150$. The direction cosines of the collective coordinates $E^k_{n,\alpha}$ have been represented by $N$ points, each corresponding to an individual atom, in the $3k_{\max}$ dimensional space of essential collective motions. In this space, the points that are located close to each other represent a similarity in directions of motion of the corresponding atoms. To obtain correlated domains, the $N$ points have been clustered using the nearest-neighbor technique [35]. This technique has been selected because no structural property, such as the number of domains, need to be assumed *a priori*. The only restriction employed is that the domains are assumed to contain more than 2 atoms.

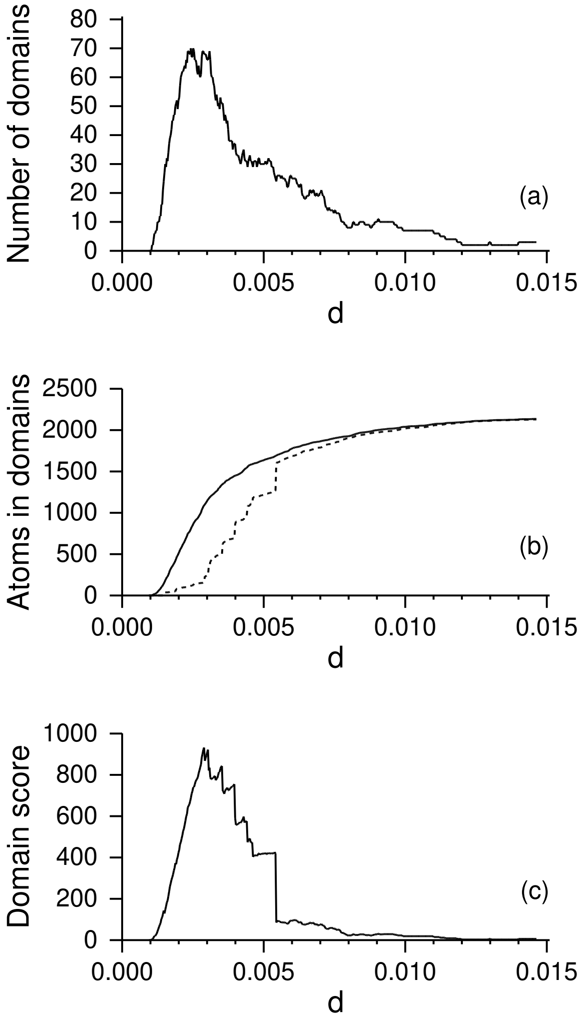



**Figure 2.** Examples of typical dependencies of the domain system in PrP 27–30 on the inter-domain distance $d$: (a) – the number of correlated domains as a function of $d$; (b) – the number of atoms in all correlated domains (solid line), and the number of atoms in the largest domain (dotted line) as functions of $d$; (c) – the difference between the number of atoms in all domains and in the largest domain as a function of $d$. Note that the distance $d$ is a dimensionless value, as it follows from the definition of the metric in the space of directional cosines of essential collective degrees of freedom.

As a part of the nearest-neighbor clustering, the inter-domain distance $d$ needs to be identified, which defines the maximum distance between the points representing atoms in the $3k_{\max}$ dimensional space of essential collective degrees of freedom, for the corresponding atoms belong to the same domain. Note that the distance $d$ is a dimensionless value, as it follows from the definition of the metric in the space of the directional cosines of the collective degrees of freedom. As it can be seen in figures 2(a) and 2(b), the identification of the domains is sensitive to the selection of $d$. Thus, no correlated domains can be identified below a minimum distance $d_{\min} \approx 0.001$, whereas most of the protein molecule is recognized as one large domain beyond a maximum distance $d_{\max} \approx 0.005 - 0.006$. The most interesting and informative breakdown of the molecule into domains is reached between these two limit values. In the next publication [34], the clustering methodology is analysed more in detail. Here examples are considered for distances $d$ that maximize the difference between the number of atoms involved in all domains and in the largest domain. In figure 2(c) it can be seen that for PrP 27–30, the corresponding optimum value of $d$ is close to 0.003.

The examples in figures 3(a) and 3(b) show five largest domains identified in PrP 27–30. The correlated domains are shown with colors, whereas atoms that do not belong to the largest domains are colored gray. An important result that emerges from the figures is that the identified correlated domains form compact groups of atoms, although the clustering formalism does not require any proximity of the locations of atoms in the primary, secondary, or tertiary structure. By the definition, a proximity in the $3k_{\max}$ dimensional space of essential collective motions reveals only a similarity in directions of the motion of atoms. The fact that this proximity identifies the compact atomic groups, confirms the viability of the clustering formalism. Another noticeable feature is that some side groups connected to the correlated domains have not been recognized as belonging to these domains. The explanation is that the motion of side groups is more flexible in comparison with the main-chain groups, which results in a weaker correlation.

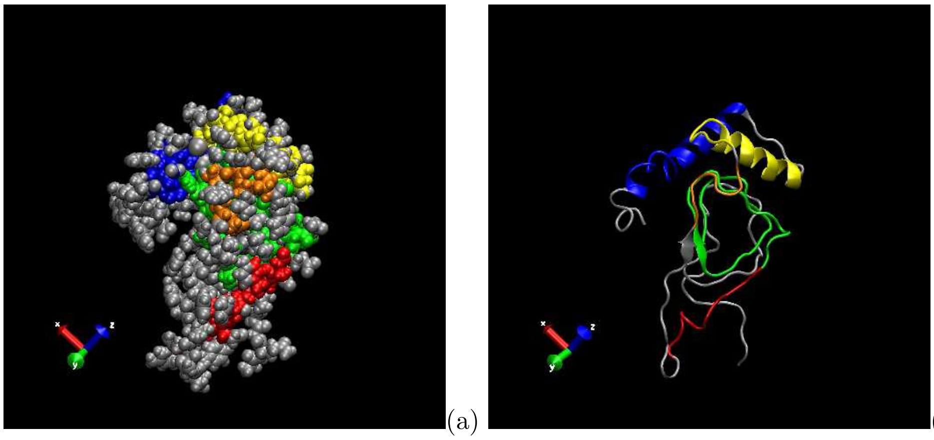
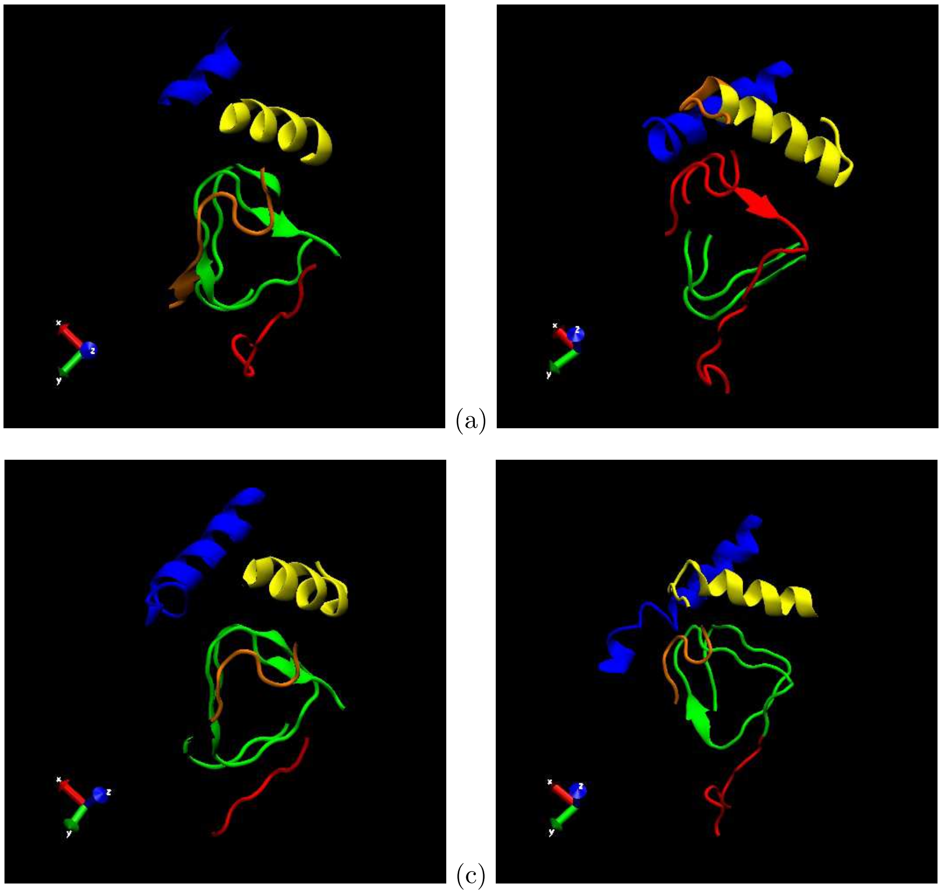

(a) (b)

**Figure 3.** Example of five largest domains identified in PrP 27–30: the general view (a) and cartoon (b). The domains are shown blue, yellow, orange, green, and red. Parts of the protein that are not involved in the domains are shown gray.

Figures 4(a–d) compare five largest domains found at various trajectory times: after equilibration (a), after 1 ns (b), after 2 ns (c), and after 3 ns (d). To obtain these, four fragments of the trajectory, each of 0.2 ns, have been analysed independently. From the figures it is evident that, despite a certain variability of the domains over the trajectory, there is a reasonable match between the secondary structure and the domains identified. Thus, significant parts of the $\alpha$-helices and $\beta$-strands are involved in correlated domains. However, the overlap between the domains and the secondary structure is not complete. Usually, only parts of $\alpha$-helices are recognized as a single domain, which reveals a flexibility within these structures. Furthermore, structure elements that are quite remotely separated in the main chain, but located near each other in the tertiary structure, are sometimes recognized as a single domain (see e.g. an example in figure 4(b)).

Another aspect that needs addressing is that the domain systems shown in figures 4(a–c) are representative of the variability of the essential collective motion. Indeed, changes in the correlated domains reflect the corresponding changes in the set of essential collective coordinates derived by PCA of different parts of the trajectory. The most important is that the changes in the domain make it possible to distinguish minor conformational variations from more significant structural changes. Thus, all four structures presented in figures 4(a–c) are somewhat different. However, not all of these differences are important. For example in figures 4(a), 4(c), and 4(d), the domains colored orange, green, and red show the same basic structure, in spite of a variability in the size and position of the domains. In figure 4(b), however, the domains have composed a different configuration, which can be interpreted as a conformational transition in the $\beta$-rich part of the protein.

One can conclude that the introduced formalism of domain identification offers a straightforward



(a) (b) (c) (d)

**Figure 4.** Five largest correlated domains in PrP 27–30 identified at various trajectory times: after equilibration (a), after 1 ns (b), after 2 ns (c), and after 3 ns (d). Off-domain parts of the protein are not shown. The examples from figures 3(a) and 3(b) correspond to (d).

and efficient methodology of characterizing the conformation stability over a molecular dynamics trajectory. If the configuration of major correlated domains does not significantly depend on the sampling interval, then the conformation space of the proteins can be considered as largely stable. Otherwise, any significant changes in the conformation space generate easily detectable changes in the domains.

## 4. Towards coarse-grained dynamics of correlated domains

After the correlated domains are identified, coarse-grained dynamics of a protein can be explicitly represented by considering the motion of the domains as a result of their interaction with each other and with the environment. Thus, each domain can be characterized by the number of atoms involved $N^{\delta}$, the mass $\tilde{M}^{\delta}$ of the domain, and the coordinates of the center of mass $\tilde{X}_{\alpha}^{\delta}$ :

$$\tilde{M}^{\delta} = \sum_{n \in \{N^{\delta}\}} m_n \, , \tag{32}$$

$$\tilde{X}_{\alpha}^{\delta}(t) = \frac{1}{\tilde{M}^{\delta}} \sum_{n \in \{N_{\alpha}^{\delta}\}} m_n X_n^{\mathrm{E}}(t) \, , \qquad \delta = 1, 2, \ldots \delta_{\max} \, , \qquad \alpha = 1, 2, 3. \tag{33}$$

Here $\delta$ denotes the domains, $\alpha$ denotes the Cartesian coordinates of the domains x, y, or z, and $X_n^{\mathrm{E}}$ represents the coordinates of individual atoms. The expression $n \in \left\{N_{\alpha}^{\delta}\right\}$ says that only $\alpha$-

th coordinates of atoms involved in the domain $\delta$ are accounted for to define the center-of-mass coordinate $\tilde{X}_{\alpha}^{\delta}$.

The coordinates of the centers of the domains of masses $\tilde{X}_{\alpha}^{\delta}$ can also be expressed through the essential collective coordinates $x^k$,

$$\tilde{X}_{\alpha}^{\delta}(t) = \sum_{k=1}^{k_{\max}} T_{\alpha\delta,k} x^k(t), \tag{34}$$

where the identity

$$X_n^{\mathrm{E}} = \sum_{l=1}^{k_{\max}} E_n^k x^k$$

is taken into account, and

$$T_{\delta\alpha,k} = \frac{1}{\tilde{M}^{\delta}} \sum_{n\in\{N_{\alpha}^{\delta}\}} m_n E_n^k .$$

Double differentiation of equation (34) over time and replacing of $\ddot{x}^k$ with equation (20) leads to

$$\ddot{\tilde{X}}_{\alpha}^{\xi} = -\sum_{k,l=1}^{k_{\max}} T_{\delta\alpha,k}\mu_{kl}^{-1}\frac{\partial U}{\partial x^l} - \sum_{k,l=1}^{k_{\max}} T_{\delta\alpha,k}\int \xi_{kl}(t-\tau)\,\dot{x}^l(\tau)\,\mathrm{d}\tau + \sum_{k=1}^{k_{\max}} T_{\delta\alpha,k} r^k(t). \tag{35}$$

Equation (35) is a formal equation of motion for the coarse-grained degrees of freedom represented by the coordinates of the centers of the domains of masses $\tilde{X}_{\alpha}^{\delta}$. If the total number of such coarse-grained degrees of freedom is equal to the number of the essential collective coordinates, $x^k$ can be expressed through $\tilde{X}_{\alpha}^{\delta}$ by $x^k = \sum\limits_{s=1}^{k_{\max}} T_{ks}^{-1}\tilde{X}^s$, where the index $s = 1, 2, \ldots k_{\max}$ replaces the pair $\{\delta\alpha\}$. This change of variables converts equation (35) into the generalized Langevin equation for the coarse-grained degrees of freedom $\tilde{X}^s$,

$$\ddot{\tilde{X}}^s = -\sum_{l=1}^{k_{\max}} V_{sl}\frac{\partial U}{\partial \tilde{X}^l} - \sum_{l=1}^{k_{\max}} \int \zeta_{sl}(t-\tau)\,\dot{\tilde{X}}^l(\tau)\,\mathrm{d}\tau + \rho^s(t), \tag{36}$$

where

$$\begin{aligned} V_{sl} &= \sum_{p,q=1}^{k_{\max}} T_{sp}\mu_{pq}^{-1}T_{lq}, \\ \zeta_{sl}(t) &= \sum_{p,q=1}^{k_{\max}} T_{sp}\xi_{pq}(t)\,T_{ql}^{-1}, \\ \rho^s(t) &= \sum_{l=1}^{k_{\max}} T_{sl}r^l(t). \end{aligned} \tag{37}$$

Equation (36) describes the coarse-grained dynamics in a protein through a few interacting domains embedded in a dissipative medium. The equation can be entirely parameterized based on the dynamics of essential collective motions discussed in section 2. If the effective masses, mean forces, and memory kernels are available for a set of essential collective coordinates, then the corresponding parameters $\partial U \big/ \partial \tilde{X}^s$, $V_{sl}$, and $\zeta_{sl}(t)$ can be identified, provided that the domains of atoms with strongly correlated degrees of freedom exist in the molecule.

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 only one collective coordinate is considered ($k_{\max} = 1$), then no more than one coarse-grained degree of freedom can be described by equation (36). If $k_{\max} = 3$, the corresponding set of 3 equations of motion can represent the $\{x, y, z\}$ coordinates of one domain, or it can describe particular degrees of freedom belonging to two or three different domains. In a general case, the number of essential collective coordinates $k_{\max}$ cannot be less than the number of the coarse-grained degrees of freedom that is necessary to describe.

The requirement to the set of essential collective coordinates to be equal or exceed the number of coarse-grained degrees of freedom that need to be described, complements the standard multivariate analysis of molecular dynamic trajectories, which only ranks the collective coordinates according to the associated mean-square displacements and does not identify what exactly the set of essential coordinates should be. The required minimum number of coarse-grained degrees of freedom provides a guidance for this choice.

Another issue of the standard ranking of the collective coordinates according to the associated mean-square displacements is that such a ranking does not define what the "sufficient" value of the displacement is for a coordinate to be essential. This implies that the standard eigenvalue ranking needs to be complemented by another criterion based on the dynamics of the essential motions. To find such a criterion, let us recall that the formalism presented in section 2 assumes that (i) the displacements related to the essential motions are significantly larger than the fluctuations; and (ii) the essential motions are significantly slower than the fluctuations. Evidently, the selection of a set of essential collective coordinates should be consistent with both assumptions. However, only the requirement (i) is fulfilled by the standard ranking of the collective coordinates $x^k$ according to the respective eigenvalues. A solution that emerges from the analysis in section 2 is to complement the ranking of eigenvalues by a comparison of the decay times $\tau_{xx}$ and $\tau_{\xi}$, which correspond to the autocorrelation function $\left\langle x^k(t)\, x^k(0)\right\rangle$ and to the memory kernel, $\xi_{kk}(t)$, respectively, for each of the potentially essential coordinates $x^k$. Indeed, the requirement that motion along the collective coordinates is slow compared to fluctuations means that $\tau_{xx}$ should be larger than $\tau_{\xi}$. Thus, the requirement $\tau_{xx} > \tau_{\xi}$ employed together with the standard ranking of the eigenvalues of the covariance matrix would provide a sufficient criterion for identifying the sets of essential collective coordinates. A potential problem may be presented by very slow modes that are not captured well enough because their characteristic time is larger than the trajectory time [16,36]. These undersampled modes combine small eigenvalues $\sigma^k$ with large decay times $\tau_{xx}$, and thus do not match neither the category of essential modes, nor the category of fluctuations. A solution is to identify and eliminate the undersampled modes, for example, through conventional drift reduction techniques. This is a natural limitation of any analysis based on molecular dynamics trajectories.

As it follows from the previous discussion, the memory kernel matrix $\xi(t)$ is one of the central quantities in the present theory. The diagonal elements $\xi_{kk}(t)$ are required for complementary ranking of the essential degrees of freedom, and the entire matrix $\xi(t)$ is needed to parameterize the coarse-grained equations of motion. For a single essential coordinate, it has been suggested to extract $\xi(t)$ from molecular dynamics trajectories by solving the memory equation for the velocity autocorrelation function, $\langle \dot{x}(t)\, \dot{x}(0)\rangle$, which is derived from the generalized Langevin equation through the well-known procedure [10,16,37]. In the case of multiple collective coordinates that is considered in this paper, the procedure analogous to that described in reference [37] leads to the following set of integral equations for $\xi_{kl}(t)$:

$$\left\langle \dot{x}^k(t)\, \dot{x}^k(0)\right\rangle = -\sum_{l=1}^{k_{\max}} \mu_{kl}^{-1} \left\langle \frac{\partial U}{\partial x^l} x^k(0)\right\rangle - \sum_{l=1}^{k_{\max}} \int \xi_{kl}(t-\tau) \left\langle \dot{x}^l(\tau)\, x^k(0)\right\rangle \mathrm{d}\tau. \tag{38}$$

The required correlation functions can be evaluated numerically from molecular dynamics trajectories [38], and the set of integral equations (38) can then be solved with respect to $\xi_{kl}(t)$.

Identification of the mean forces $-\partial U/\partial x^k$ is the most challenging implication of the theory. It has been suggested in the literature to evaluate the potential of mean force $U(x)$ from the phase space density $\Psi(x)$ that is derived from molecular dynamics trajectories [7,14,16],

$$U(x) = -\beta^{-1} \log(\Psi(x)), \tag{39}$$

provided that the set of snapshots obtained from molecular dynamics simulation satisfies the condition of ergodicity [7,14,16]. In the case of multiple essential coordinates, however, the mean force along a particular collective coordinate $x^k$ is a function of other coordinates. The expression for the mean force $-\partial U\left(x^1, x^2, \ldots x^{k_{\max}}\right)/\partial x^k$, in principle, can be derived analytically based on the formalism from sections 2.2 and 2.3. This would also make it possible to predict the variability of domains with time. To accomplish this, however, a self-consistent solution of equations (7) and (20) needs to be found. Solving this fundamental challenge appears to be one of the most important and promising milestones in the future development of the theory. At the present stage, however, the major purpose of the theory is characterization and comparison of protein conformations based on molecular dynamics trajectories. For this particular purpose, employing dependencies analogous to equation (39)in order to evaluate $U(x)$ projected onto particular collective coordinates [14] appears to be a reasonable approximate solution, provided that the output of the molecular dynamics trajectory satisfies the requirements for such an interpretation [7,14,16].

## 5. Summary

This work introduces a comprehensive theoretical methodology of describing the conformational dynamics of proteins based on the projection operator method [22–25]. The essential collective degrees of freedom defined by the principal component analysis of the molecular dynamics trajectory are used as dynamic variables in the formalism. The explicit form of the corresponding projection operator is obtained, and the projection technique is employed to derive systems of the coupled generalized Langevin equations for both individual atomic degrees of freedom and essential collective degrees of freedom. The number of the essential degrees of freedom is not limited in the theory. Unlike other studies of protein dynamics, the present theory is valid for any number of essential coordinates. In particular, the coupling of relevant dynamic variables is explicitly included in the equations of motion. The theory includes the model with a single essential degree of freedom, as a particular case.

Based on the analysis of the coupling of the dynamic variables, a consistent definition of correlated domains in a protein has been introduced. The domains, which are supposed to serve as major building blocks for coarse-grained modelling of proteins, are defined as groups of atoms whose Cartesian coordinates show a strong coupling in the generalized Langevin equation of motion. For such groups of atoms, the direction cosines of the essential collective degrees of freedom adopt similar values. Accordingly, the domains are identified through a simple clustering procedure. Unlike the existing approaches to the identification of the domains in proteins, subject to clustering are the directional cosines of the essential collective degrees of freedom in the phase space, and not translations and/or rotations of individual atomic groups, which makes the formalism immune to noises. Furthermore, no limiting assumptions are made regarding the structure of domains, their number, or interatomic interactions in the protein. The formalism of domain identification is general, physically transparent, and intimately related to the essential dynamics of the protein.

An example of identification of correlated domains is provided for the misfolded isoform of a prion protein, PrP 27–30 [33]. 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 domains have also shown a reasonable match with the secondary structure; however, there is no complete similarity. Some domains follow closely particular elements of secondary structure or their parts, while others are composed of different elements that are located near each other in the tertiary structure but are quire remotely separated in the main chain.

It has been demonstrated that the introduced formalism of domain identification offers a straightforward and efficient methodology of characterizing the conformation stability of proteins over a molecular dynamic trajectory. If the configuration of major correlated domains does not significantly depend on the sampling interval, then the conformation of the proteins can be considered as largely stable. Otherwise, changes in the essential collective coordinates cause easily detectable changes in the correlated domains.

The potential of building analytic coarse-grained models describing conformational motions in a protein through a few interacting domains embedded in a dissipative medium has been analysed in detail. For the first time, generalized Langevin equations of motion for the Cartesian coordinates of the correlated domains are derived and parameterized analytically based on the equations of motion for the essential collective coordinates. It is demonstrated that the number of addressable coarse-grained degrees of freedom cannot exceed the number of essential collective coordinates identified in the macromolecule. Dynamic criteria for identifying the set of essential collective coordinates are identified. Methodologies of parametrizing the equations of motion for the correlated domains are outlined, and potential further developments are discussed. Thus, a fundamental challenge and one of the most important and promising future milestones in the extension of the formalism is prediction of the dynamic variability of the energy landscape that generates changes in the domain structure of the protein.

## Acknowledgements

The author thanks A. Kitao, A. Kobrin, A. Potapov, and A. Kovalenko for their discussion of the work and helpful advice, H. Wille for provision of the structure of PrP 27–30 [33], and T. Yamazaki for generating the molecular dynamics trajectory for the analysis [34]. The molecular structure images were generated with VMD 1.8.3.

## References

1. Kitao A., Hirata F., Go N., Chem. Phys., 1991, **158**, 447.
2. Garcia A.E., Phys. Rev. Lett., 1992, **68**, 2696.
3. Hayward S., Kitao A., Hirata F., Go N., J. Mol. Biol., 1993, **234**, 1207.
4. Amadei A., Linssen A.B.M., Berendsen H.J.C., Proteins, 1993, **17**, 412.
5. Van Aalten D.M.F., De Groot B.L., Findlay J.B.C., Berendsen H.J.C., Amadei A., J. Comput. Chem., 1997, **18**, 169.
6. Kitao A., Go N., Curr. Opin. Struct. Biol., 1999, **9**, 174.
7. Grubmuller H., Phys. Rev. E, 1995, **52**, 2893.
8. Amadei A., de Groot B.L., Ceruso M.-A., Paci M., Di Mola A., Berendsen H.J.C., Proteins, 1999, **35**, 283.
9. Eastman P., Pellegrini M., Doniach S., J. Chem. Phys., 1999, **110**, 10141.
10. Sagnella D.E., Straub J.E., Thirumalai D., J. Chem. Phys., 2000, **113**, 7.
11. Ansari A., J. Chem. Phys., 2000, **112**, 5.
12. Oliva B., Daura X., Querol E., Aviles F.X., Tapia O., Theor. Chem. Acc., 2000, **105**, 101.
13. Bu L., Straub J.E., Biophysical Journal, 2003, **85**, 1429.
14. Kosztin I., Barz B., Janosi L., J. Chem. Phys., 2006, **124**, 064106.
15. Moritsugu K., Smith J.C., J. Phys. Chem., 2006, **110**, 5807.
16. Lange O.F., Grubmuller H., J. Chem. Phys., 2006, **124**, 214903.
17. Yesylevsky S.O., Kharkyanen V.N., Demchenko A.P., Biophys. J., 2006, **91**, 670.
18. Hayward S., Kitao A., Berendsen H., Proteins, 1997, **27**, 425.
19. Hayward S., Berendsen H., Proteins, 1998, **30**, 144.
20. Hinsen K., Proteins, 1998, **33**, 417.
21. Hinsen K., Thomas A., Field M.J., Proteins, 1999, **34**, 369.
22. Mori H., Prog. Theor. Phys., 1965, **33**, 423.
23. Mori H., Prog. Theor. Phys., 1965 **34**, 399.
24. Kob W., Supercooled liquids, the glass transition, computer simulations, 2002 Lecture Notes for Les Houches 2002 Summer School – Session LXXVII: Slow Relaxations, Nonequilibrium Dynamics in Condensed Matter, J.-L. Barrat, M. Feigelman, J. Kurchan, Jean Dalibard eds., EDP Sciences; Springer-Verlag, Paris, 2003, p. 201-269.
25. Balucani U., Lee M.H., Tognetti V., Phys. Rep., 2003, **373**, 409.
26. Magalinski V.B., Sov. Phys. JETP, 1959, **9**, 1381.
27. Cortes E., West B. J., Lindenberg K., J. Chem. Phys., 1985, **82**, 15.
28. Zhou H.-X., Zwanzig R., J. Phys. Chem. A, 2002, **106**, 7562.
29. Hanggi P., Talkner P., Borkovec M., Rev. Mod. Phys., 1990, **62**, 251.
30. Ingold G.-L., Lecture Notes in Physics, A. Buchleiter, K. Hornberher Eds., LNP 611, Springer, Berlin, 2002, 1-53.
31. Hanggi P., Ingold G.-L., Chaos, 2005, **15**, 026105.

32. The degenerated case of $E^k_{n,\alpha} = 0$ for all $k$ means that dynamics of such a "domain" is not captured by the set of essential collective degrees of freedom identified, and thus this case is irrelevant to the theory.
33. Govaerts C., Wille H., Prusiner S.B., Cohen F.E., Proc. Natl. Acad. Sci. USA, 2004, **101**, 8342.
34. Stepanova M., Berjanskii M., Wishart D., Yamazaki T., Kovalenko A. More details about the molecular dynamics simulation with examples of applications of the theory are given in the forthcoming publication [in preparation].
35. Jain A.K., Murty M.N., Flynn P.J., ACM Computing Surveys, 1999, **3**, 264.
36. Balsera M.A., Wriggers W., Oono Y., Schulten K., J. Phys. Chem., 1996, **100**, 2567.
37. Berkowitz M., Morgan J.D., Kouri D.J., McCammon J.A., J. Chem. Phys., 1981, **75**, 2462.
38. Xing C., Andricioaei I., J. Chem. Phys., 2006, **124**, 034110.

# До крупнозернистого моделювання протеїнів

М.Степанова

Національний інститут нанотехнології, Національна дослідницька рада Канади, факультет електричної і комп'ютерної інженерії, університет м. Альберта, Саскачеван драйв, Едмонтон, Альберта, Канада

Отримано 31 травня 2007 р.

Ця стаття вводить базисну теоретичну основу для опису конформаційної динаміки протеїнів через систему взаємодіючих доменів. Суттєві колективні ступені вільності, отримані аналізом провідної компоненти траєкторії молекулярної динаміки, використовуються як динамічні змінні з допомогою техніки проекційного оператора, що окреслює запропонований формалізм. Отримано явну форму відповідного проекційного оператора. Проекційний метод застосований для встановлення системи зв'язаних узагальнених рівнянь Ланжевена для індивідуальних атомних ступенів вільності та суттєвих колективних ступенів вільності в протеїні. Визначення кореляційних доменів в протеїнах вводиться на основі аналізу суттєвої динаміки. Наведені приклади ідентифікації таких доменів. Система зв'язаних узагальнених рівнянь Ланжевена була отримана, шляхом представлення протеїну через кілька взаємодіючих областей, поміщених в дисипативне середовище. Обговорюються подальший розвиток і можливості застосування запропонованого формалізму.

**Ключові слова:** *динаміка протеїнів, конформаційні зміни, теорія і моделювання, проекційний оператор, аналіз провідної компоненти*

**PACS:** *87.15.He, 05.10.Gg, 87.15.Aa, 02.50.Sk*