Identification of Dynamic Structural Domains in Proteins, Analysis of Local Bond Flexibility, and Application for Interpretation of NMR Experiments

M. Stepanova

National Institute for Nanotechnology, National Research Council of Canada;
11421 Saskatchewan Drive, Edmonton, Alberta, T6G 2V4, Canada
Phone: 1-780-641-1717, e-mail: Maria.Stepanova@nrc-cnrc.gc.ca

ABSTRACT

A novel theoretical methodology is described that allows identifying dynamic structural domains and analyzing local flexibility in proteins. The methodology employs a multiscale approach that combines definition of essential collective coordinates based on the covariance analysis of molecular dynamics trajectories, construction of the Mori projection operator with these essential coordinates, and analysis of the corresponding generalized Langevin equations. The domains are associated with relatively stable and regions in the protein, whereas off-domain regions are relatively soft. The applications include the domain coarse-graining and characterization of the local flexibility in protein G and prion proteins. The results are compared with published NMR experiments. The methodology is apt to provide rigorous dynamic scores and characterization tools for structural biology, bioinformatics, and rational drug design.

Keywords: protein conformations, multiscale modeling, generalized Langevin dynamics, interpretation of NMR measurements.

1 INTRODUCTION

Multiscale coarse-graining methodologies for proteins are of keen importance for basic understanding of their structure and function, as well as for development of efficient molecular simulation protocols, bioinformatics algorithms, and scores for *in-silico* aided biomolecular engineering [1,2]. However, identification of coarse-grained units in proteins is a tremendously complex task because of the complex hierarchical structure of proteins, absence of periodicity, and their ability to change the spatial conformation. Over the last decade there was a significant effort to develop coarse-grained models of proteins based on their molecular dynamics (MD) trajectories [2-4]. Difficulties arise even with the very definition of the coarse-grained units, which sometimes include rather vague criteria such as being a visually recognizable substructure in the protein [4]. In the most elaborate approaches [5-8], the coarse-grained 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 priory. Moreover, the differences in motion that need to be captured are very subtle and susceptible to uncertainties (e.g. high-frequency noise, sampling scheme, etc.,) and thus a proper filtering is required, which is complex and computationally expensive. Results of domain identification depend on the assumptions made and particular techniques employed, which vary significantly in different publications [4].

This paper presents a universal and dynamically justified methodology for identification of coarse-grained domains in proteins that has been introduced recently [9]. The methodology employs a multiscale approach combining the identification of essential collective coordinates based on the covariance analysis of molecular dynamics trajectories, construction of the Mori projection operator with these essential coordinates, and analysis of the corresponding generalized Langevin equations (GLE). The structural domains are identified as groups of atoms whose motions show a dynamic coupling in the GLE formalism. The methodology is independent on any reference configurations, is compatible with any model of interatomic interactions in the macromolecule, and generically immune to short-wavelength noises because the domains are identified in the space of essential collective coordinates. Because the approach is based on a rigorous theory, the outcomes are physically transparent: the dynamic structural domains are associated with regions of relative rigidity in the protein, whereas off-domain regions are relatively soft. This allows introducing a numeric descriptor for the local flexibility [10], which can be identified with atomic-level resolution. In the present contribution, the background of the methodology is outlined with the examples of applications for medium-size proteins such as protein G and prion proteins.

2 METHODS

As the input, a short (~0.1 ns) MD trajectory of a solvated protein is used. From this trajectory, the collective coordinates $\vec{E}^k = \{E_1^k, E_2^k, ...E_{3N}^k\}$, k= 1,2,…3N are identified by the principal component analysis (PCA) [11]. Here N is the number of atoms in the system, 3N is the corresponding number of the Cartesian coordinates of atoms, and $\vec{E}^k$ are the normalized eigenvectors of the covariance matrix for the trajectory, which can be described by a set of time-dependent Cartesian coordinates, $\vec{X}(t) = \{X_1(t), X_2(t), ...X_{3N}(t)\}$. One can consider the eigenvectors $\vec{E}^k$ as the intrinsic coordinate frame in the configuration space, and project on them the trajectory, $\left(\vec{X}(t)\cdot\vec{E}^k\right) = \sum_{i=1}^{3N} E_i^k X_i(t) = x^k(t)$, k = 1,2,…3N. In these equations, the number of atoms N may be of the order of 1000 or more. However, it is possible to rank the projections x^k according to the corresponding mean square displacements, and consider a truncated set of collective coordinates, $\vec{E}^1, \vec{E}^2, ...\vec{E}^{k_{max}}$, which include only those coordinates that correspond to the highest magnitude of the displacements [9,11]. In practice, the number of coordinates k_{max}=10–30 is often sufficient. This truncated set of collective coordinates is sometimes referred to as the essential degrees of freedom. The complementary set of collective coordinates, $\vec{E}^{k_{max}+1}, \vec{E}^{k_{max}+2}, ...\vec{E}^{3N}$ can be attributed to fluctuations.

As the next step, the Mori projection operator P and its complement 1–P [12] are constructed using the sets $\vec{E}^1, \vec{E}^2, ...\vec{E}^{k_{max}}$ and $\vec{E}^{k_{max}+1}, \vec{E}^{k_{max}+2}, ...\vec{E}^{3N}$, respectively, such that [9]

$$P\vec{X}(t) = \vec{X}^E(t), \qquad (1-P)\vec{X}(t) = \vec{X}^{1-E}(t). \qquad (1)$$

Here the vector $\vec{X}(t)$ represents the Cartesian coordinates of all atoms as a function of time, and $\vec{X}^E(t)$ and $\vec{X}^{1-E}(t)$ represent the essential component and the fluctuations, respectively. Clearly, $\vec{X} = \vec{X}^E + \vec{X}^{1-E}$. Next, using the technique described in Ref. [9], the sets of generalized Langevin equations are derived for both the projections x^k (k = 1,2,… k_{max}) and the projected atomic coordinates X_i^E (i = 1,2,…3N). The latter can be represented by

$$\ddot{X}_i^E(t) = \qquad (2)$$
$$= -\sum_{j=1}^{3N} C_{ij} m_j^{-1} \left[\frac{\partial U}{\partial X_j^E} - \sum_{l=1}^{3N} \int_o^t Z_{lj}(t-\tau)\dot{X}_j^E(\tau)d\tau + \rho_i(t) \right]$$

where U is the potential of mean force, $Z_{ij}(t)$ is the memory kernel, and $\rho_i(t)$ is the random force.

The coefficients C_{ij} in Eq.(2) represent the dynamic connection (coupling) of the atomic coordinates i and j,

$$C_{ij} = \sum_{k=1}^{k_{max}} E_i^k E_j^k , \qquad (3)$$

where E_i^k and E_j^k are the directional cosines of the essential eigenvector $\vec{E}^k$ in the 3N-dimensional configuration space. It can be shown [9] that the condition of a strong dynamic coupling between the coordinates i and j is equivalent to the requirement that the corresponding direction cosines are equal or close in magnitude, $E_i^k \approx E_j^k$.

Accordingly, the coarse-grained dynamic domains can be identified as groups of atoms, for which the direction cosines of the essential collective degrees of freedom E_i^k adopt similar values for each k. More simply speaking, the domains represent regions of a relative rigidity in the protein, whereas off-domain regions are relatively soft. It should be noted that in contrast to other studies [5-8], no assumption regarding any elementary building blocks and/or interatomic interactions has been made to define the dynamic domains [9].

Once the domains of correlated motion are defined, it is possible to apply the methodology to identify coarse-grained subunits in particular proteins. The results considered in this work employ 0.2-ns MD trajectories, using GROMACS 3.2.1 with the GROMOS96 force field, for protein G and prion proteins [9,10]. For essential collective coordinates, 10 to 30 principal components with the highest eigenvalues were selected. The direction cosines E_i^k were represented by N points, 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 directions of motion of the corresponding atoms. To obtain correlated domains, the N points have been clustered using the nearest-neighbor technique [9].

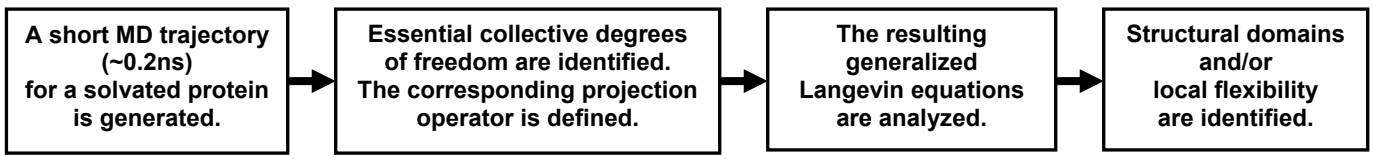

Figure 1: Outline of the identification of dynamic structural domains and local flexibility in macromolecules.

NSTI-Nanotech 2009, www.nsti.org, ISBN 978-1-4398-1783-4 Vol. 2, 2009

In addition to the identification of dynamic domains, the same theoretic framework can be employed to characterize the local flexibility of inter-molecular bonds in a protein. The corresponding descriptor for the local flexibility, F, is given by [10]

$$F(i) = \sum_{k=1}^{3k_{max}} \left(E_i^k - \varepsilon^k\right)^2 , \qquad (4)$$

where i denotes the atoms for which the local flexibility is identified, and $\varepsilon^k = \frac{1}{n}\sum_{j=1}^{n} E_j^k$ describes the motion of the entire molecule. The methodology allows identifying the flexibility for every atom, in which case n is equal to the total number of atoms in the molecule. For comparison with experiments, however, it is often more convenient to analyze the flexibility for α-carbon atoms in the main chain. In this case, the sum in the expression for ε^k is taken over the coordinates of C_α atoms [10]. A low value for F indicates a strong correlation with motion of the entire molecule (i.e. a low flexibility), whereas high F values identify more mobile and/or flexible locations.

The methodology for the dynamic coarse-graining and characterization of the local flexibility in proteins is outlined in Figure 1.

3 EXAMPLES OF APPLICATIONS

The examples in Figure 2 show three largest domains identified for the fragment B1 of protein G [9]. The correlated domains are shown with colors, whereas atoms that do not belong to the largest domains are shown in white. An important result that emerges from the figures, is that the identified correlated domains form compact groups of atoms, although the model employed 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 the atom's motion. The fact that these correlations identify compact atomic groups confirms the validity of the clustering formalism. Another noticeable feature is that some side groups connected to the correlated domains have not been recognized as a part of these domains (see Figure 2(a)). The explanation is that the side groups' motion has more flexibility in comparison to the main-chain groups, which results in a weaker correlation.

From Figure 2(b) it can be seen that the identified domains also show a reasonable match with the secondary structure in protein G; however, there is no complete similarity. Some domains follow closely particular elements of secondary structure or their parts, and others are composed of different elements that are located near each other in the tertiary structure but are quite remotely separated in the main chain.

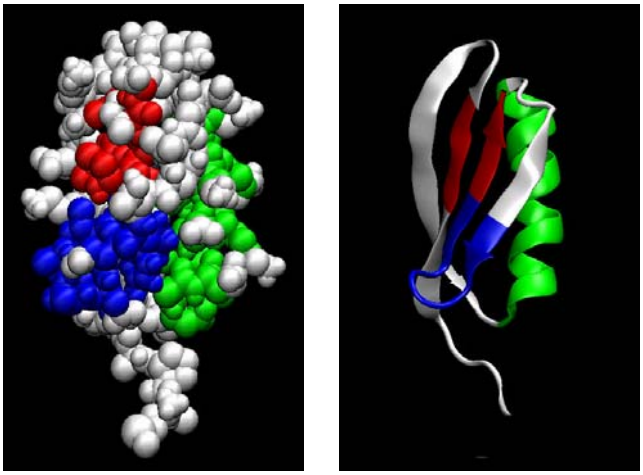

Figure 2: Example of dynamic domains for protein G [9]. Three largest domains are shown by colors. Off-domain regions are white. The figures were generated using the VMD software [13].

Comparison with the model-free S^2 order parameters derived from NMR experiments [14,15] has shown that the predicted large domains match the locations of high S^2 levels (see Table 1). Furthermore, in addition to the structural coarse-graining, the local flexibility in a molecule can be characterized by the descriptor given in Eq. (4). In Figure 3, the normalized F profile for human prion protein is compared with the corresponding RCI dependencies derived from the NMR chemical shifts [16] as reported elsewhere [10]. The predicted positions of major maxima and minima in Figure 3 match remarkably well with the NMR results. Although there is a difference in the relative heights of individual maxima, which arises from a slight

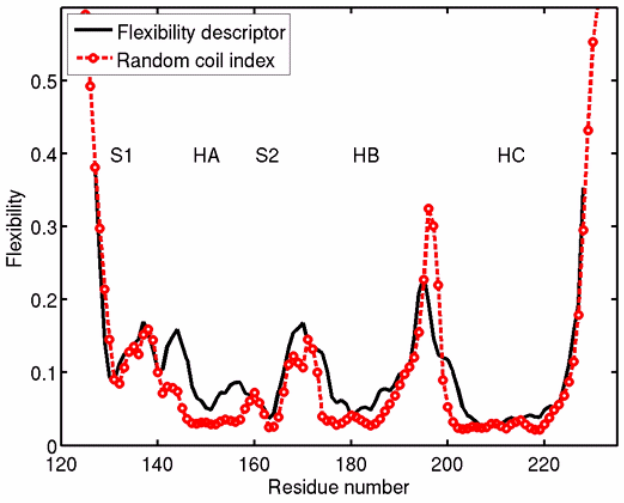

Figure 3: The normalized main-chain flexibility descriptor F (solid line) and the NMR-derived random coil index (circles) for human prion protein [10].

Table 1: Comparison of the numeric results with NMR experiments and examples of the interpretation.

Protein	Model Prediction	Comments	Ref.
Protein G	Dynamic domains	Predicted dynamic domains identify the locations of relative rigidity, which correspond to high levels of the model-fee order parameters derived from NMR measurements [14,15].	[9]
Protein G	Analysis of dynamic topology	Special dynamic topology analysis explains the cross-strand correlations observed by the residual dipolar coupling [15].	[9]
Human PrP	Dynamic domains	Predicted dynamic domains indicate the locations of low flexibility corresponding to low levels of RCI derived from NMR data [16].	[10]
Chicken PrP	Dynamic domains	4 out of 5 major dynamic domains correspond to low levels of the RCI derived from NMR chemical shifts [17]. The 5^{th} domain, located in the loop between the helices HB and HC, seems to be relatively rigid but dynamically uncoupled with the rest of the molecule.	[10]
Human PrP	Main-chain flexibility	The positions of major maxima and minima of the numeric flexibility descriptor match the RCI profiles derived from NMR experiments [16], confirming the variations of the flexibility along the main chain.	[10]

difference in the physical meaning of the *F* parameter and the RCI, the numeric F-profile shows an excellent match with the locations of relative rigidity and softness determined from experiments.

Table 1 gives more examples of comparison of the numeric results with experiments and their application to interpret NMR structural data.

4 SUMMARY

A novel numeric methodology has been developed for dynamic coarse-graining of proteins, based on a rigorous dynamic theory connecting the collective coordinates in the macromolecule with its dynamic structural properties. The latter include the coarse-grained domains and the local flexibility in proteins, which are determined employing MD trajectories as the input. The numeric predictions match reasonably NMR-derived structural data. Remarkably, the numeric predictions derived from short (less than a nanosecond) trajectories agree with NMR data obtained over significantly longer times. This indicates that the methodology is capable to capture stable structural trends that may persist over longer times than the MD trajectories from which they were derived.

The methodology provides a natural criterion for characterization and comparison of the stability of macromolecules and their complexes. Further applications may include interpretation of NMR structural data and rectification of protein's tertiary structures. The analysis of the stability of protein-ligand binding, which is an important component of the rational drug design, is another promising line of applications. Finally, because the methodology of coarse-graining is based on a rigorous theoretical background, it also can be employed as a starting point to develop new mesoscopic models of proteins, as well as improve molecular dynamics protocols and bioinformatics algorithms.

REFERENCES

[1] P. Aloy and R.B. Russell, Nature Reviews, Molecular Cell Biology, 7, 188, 2006.
[2] V. Tozzini, Curr. Opin. Struct. Biol., 15, 144, 2005.
[3] I. Bahtar and A.J. Rader, Curr. Opin. Struct. Biol., 15, 585, 2005.
[4] S.O. Yesylevsky, V.N. Kharkyanen, and A.P. Demchen-ko, Biophys. J., 91, 670, 2006.
[5] S. Hayward, A. Kitao, and H. Berendsen, Proteins, 27, 425, 1997.
[6] S. Hayward and H. Berendsen, Proteins, 30, 144 1998.
[7] K. Hinsen, Proteins, 33, 417, 1998.
[8] K. Hinsen, A. Thomas, and M.J. Field, Proteins, 34, 369, 1999.
[9] M. Stepanova, Phys. Rev. E 76, 051918, 2007.
[10] N.Blinov, M.Berjanskii, D.Wishart, and M.Stepanova, Biochemistry, 48, 1488, 2009.
[11] A. Kitao, F. Hirata, and N. Go, Chem. Phys. 158, 447, 1991.
[12] H. Mori, Prog. Theor. Phys. 33, 423, 1965, ibid., 34, 399, 1965.
[13] Humphrey, W., Dalke, A., and Schulten, K, J. Mol. Graph. 14, 33, 1996.
[14] M. J. Stone, S. Gupta, N. Snyder, and L. Regan, J. Am. Chem. Soc. 123, 185, 2001.
[15] G. Bouvignies, P. Bernado, S. Meier, K. Cho, S. Grzesiek, R. Brüschweiler, and M. Blackledge, PNAS 102, 13885, 2005.
[16] R. Zahn, A. Liu, T. Luhrs, et.al. Proc. Natl. Acad. Sci. U. S. A. 97, 145, 2000.
[17] L. Calzolai, D.A. Lysek, D.R. Pérez, P. Güntert, and K. Wüthrich, Proc. Natl. Acad. Sci. U.S.A. 102, 651, 2005.