Long-Range Structural Regularities and Collectivity of Folded Proteins

Canan Atilgan[1], Ibrahim Inanc[1], and Ali Rana Atilgan[1]

[1] Faculty of Engineering and Natural Sciences, Sabanci University, 34956 Istanbul, Turkey

ABSTRACT

Coarse-grained network models of proteins successfully predict equilibrium properties related to collective modes of motion. In this study, the network construction strategies and their systematic application to proteins are used to explain the role of network models in defining the collective properties of the system. The analysis is based on the radial distribution function, a newly defined angular distribution function and the spectral dimensions of a large set of globular proteins. Our analysis shows that after reaching a certain threshold for cut-off distance, network construction has negligible effect on the collective motions and the fluctuation patterns of the residues.

INTRODUCTION

Globular proteins show diversified structures and sizes, yet, it has been claimed that they display a nearly random packing of amino acids with strong local symmetry on the one hand [1], and that they are regular structures that occupy specific lattice sites, on the other [2]. It was later shown that this classification depends on the property one investigates, and that proteins display "small-world" properties, where highly ordered structures are altered with few additional links [3]. Furthermore, packing density of proteins scales uniformly with their size [4] which causes them to show similar vibrational spectral characteristics to those of solids [5,6]. Coarse grained protein models have shown a great success in the description of the residue fluctuations and the collective behavior of proteins [7]. Using a single parameter harmonic potential [8], the large amplitude motions of proteins in the native state have been predicted successfully using normal mode analysis [9]. This model, with its simplicity, speed of calculation and relying mostly on geometry and mass distribution of the protein, demonstrates that a single-parameter model can reproduce complex vibrational properties of macromolecular systems.

Following the uniform harmonic potential introduced by Tirion [10], residue level application of elastic network models paved the way for the Gaussian Network Model, which is based on the energy balance of the system at the energy minimum [9,11]. Elastic models based on the force balance around each node [12] led to the development of the so called Anisotropic Network Model (ANM) [13]. The applications of these models on many proteins show successful results in terms of predicting the collective behavior of proteins. Despite numerous applications comparing the theoretical and experimental findings on a case-by-case basis [14], only a few attempted a statistical assessment of the models. In another study where 170 pairs of structures were systematically analyzed, it was shown that the success of coarse-grained elastic network models may be improved by recognizing the rigidity of some residue clusters [15]. The results have also been shown to be protein dependent [16].

To date, the structures that form the basis of the network models have been generated from certain rules of thumb. Connectivity is assumed between C_α atom pairs in a range of 8 – 14 Å in different studies in the literature based on the argument that (i) the eigenvalue distributions obtained from the modal decomposition are similar to those obtained from the full-atom NMA description of proteins, or (ii) these provide atomic fluctuation profiles that display the largest correlation with the experimental B-factors. In this study, we use a systematic approach on a large set of globular proteins with varying architectures and sizes to find a basis for why the network models work well to define certain properties of the system. We show that the network construction is free of the cut-off distance problem once a certain baseline threshold is accessed, if one is interested in the collective motions and the fluctuation patterns of the residues.

COMPUTATIONAL DETAILS

Network construction. We base our calculations on a set of 595 proteins with sequence homology less than 25% and sizes spanning 54–1021 residues [17]. A protein of N residues is treated as a residue-based structure, where the C_α atom of each amino acid is considered as a node, and the coordinates of the protein are obtained from the protein data bank (PDB) [18]. The network information is contained in the N × N adjacency matrix, A, of inter-residue contacts, whose elements A_{ij} are taken to be 1 for contacting pairs of nodes i and j, and zero otherwise. We establish a link between two nodes if they are within a cut-off distance r_c of each other.

Radial and angular distribution functions. The radial distribution function (RDF), $g(r)$, is a measure of the correlation between the locations of particles within a system, measured as the probability of finding another particle at a distance, r, from a chosen particle, normalized by the volume element. In the current work, we are not only interested in the number distribution of particles around a given node, but also concentrate on the link structure. By treating all neighbors of a node equivalently, we find that as r_c is increased with the addition of new neighbors to each node, the resultant vector, $\mathbf{Q}_i$, on node i due to all its neighbors, j, converges to a certain location, $\mathbf{Q}_i = \sum_j A_{ij}\,\mathbf{R}_{ij}$, where $\mathbf{R}_{ij}$ is the unit vector connecting residue pairs i and j, and A_{ij} are the elements of the adjacency matrix. An example is shown on a 54 residue α-helical protein (PDB code: 1enh) in Fig. 1, where the length of a red vector is proportional to r_c and demonstrate that at small r_c, the neighbors of a node are at distinct locations, whereas with increasing r_c, the new nodes are added in a spherically symmetrical manner so that the resultant vector, $\mathbf{Q}_i$, is slightly modified. To quantify this behavior, we define the angular distribution function (ADF), which is the distribution of angular change, $\Delta\varphi$, of the resultant vector obtained from the contacting residues at a distance r to $r+dr$ to the reference residue, where dr is a perturbation on the distance r:

$$\cos\Delta\varphi_i(r) = \left(\sum_j A_{ij}\mathbf{R}_{ij}\right)_r \cdot \left(\sum_j A_{ij}\mathbf{R}_{ij}\right)_{r+dr} = \mathbf{Q}_i|_r \cdot \mathbf{Q}_i|_{r+dr} \qquad (1)$$

Anisotropic network model. In ANM**,** once the networks are formed, the interactions between connected nodes is considered to be harmonic [13], coupled by elastic springs having a uniform force constant γ. Thus, the overall potential of the molecule is given by the sum of all harmonic potentials among interacting nodes. For a network of N nodes, the Hessian is a $3N$ x $3N$ matrix whose pseudo-inverse is the covariance matrix $\mathbf{C}$ that can be expressed in terms of the $3N - 6$ non-zero eigenvalues λ_k and corresponding eigenvectors $\mathbf{u}_k$ of $\mathbf{H}$ as, $\mathbf{C} = \mathbf{U}\,\mathbf{\Lambda}^{-1}\,\mathbf{U}^{\mathrm{T}}$. The residue fluctuations are predicted by the ANM for residue i from the trace of $\mathbf{C}_{ii}$. Theoretically, they are related to the B-factors determined from X-Ray crystallografic data through the relation, $\mathbf{B}_i$ =

$(8\pi^2 k_B T/3\gamma)\, tr(\mathbf{C}_{ii})$, where k_B is the Boltzmann constant and T is the absolute temperature. The value of γ is determined *a posteriori* if experimental data are available, and does not affect the fluctuation profile of residues.

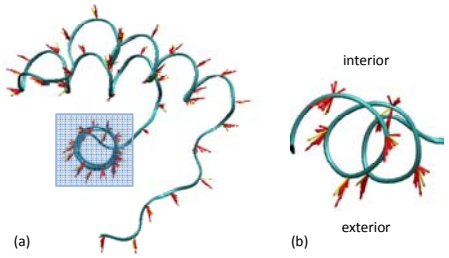

Figure 1. (a) The negative of the resultant vectors acting on the nodes, $-\mathbf{Q}_i$, exemplified by a 54 residue protein (PDB code: 1enh). The length of each red vector is proportional to r_c used, shortest at 7 Å and longest at 15 Å. (b) Part of the helix marked by the square in (a) is magnified; "exterior" refers to the solvent contacting part of the helix, and "interior" marks the side facing protein core.

RESULTS AND DISCUSSION

Structural heterogeneity of amino acid distributions in proteins. The RDF, $g(r)$, of the residues is presented in Fig. 2a for distances up 20 Å, recorded at 0.1 Å resolution. We find that the first sharp peak in $g(r)$ ends at ca. 6.7 Å corresponding to the first coordination shell, the second coordination shell occurs at 8.5 Å. Broader peaks ending at 10.5 and 12 Å are identified as the third and fourth shells. At larger distances, $g(r)$ monotonically decreases, indicating that the coarse-grained residue beads do not undergo further ordering in the liquid-like environment. In Fig. 2a we also display the ADF, $g(\varphi)$, for the same set of proteins in the same distance range. We find that the main peaks of ADF and RDF overlap, the only difference in the general character of the two distribution functions being in the third and fourth coordination shells. In RDF, we find that a similar number of particles per unit volume exist in these two coordination shells (same height in the distribution). The ADF provides the additional information that, due to the asymmetry in the intensities of the third and fourth coordination shells, these particles are clustered in relatively more ordered directions in the third shell, quantified by the increase in ADF to ca. 5°. The ADF provides the valuable information that the additional particles are taken into account as more concentric spherical shells of 0.1 Å diameter are added (recall Fig. 1), have a preferred direction of clustering at the regions of higher number density. Conversely, at larger distances, the new neighbors carry directionality that cancel each other out, as would be expected from a random packing of spheres, quantified by the monotonical decrease in $g(\varphi)$.

Since globular proteins may be considered to be made up of a core region surrounded by a molten layer of surface residues [19], it is of interest to distinguish the topological differences between the core and the surface (Fig. 2b). We observe that core residues have larger angular changes in the resultant vector, $\mathbf{Q}_i$ (Eq. 2) compared to the surface residues. Thus, the resultant vector on the surface residues rapidly converges to a given directionality specific to each residue at short distances, the additional links at higher distances arriving in directions that cancel out. The overall structural heterogeneity is detected much clearly in the $g(\varphi)$ of the core residues. However, the heterogeneity in the first coordination shell is more pronounced over that of the second for the surface residues, possibly due to the loose packing in this region.

Density of vibrational normal modes. The vibrational normal mode spectra, $g(\omega)$, of proteins was originally studied by ben-Avraham for five proteins with sizes in the range of 39 – 375

residues, the data collapsing on a single curve, especially in the slow mode region [5]. The density of states was found to increase linearly with the frequency in this region, implying a spectral dimension of d_s = 2 and deviating from the Debye model of elastic solids where the expected value is 3 [20]. The anomalous spectral dimensions of proteins was also confirmed by inelastic neutron scattering experimental measurements, which yielded $d_s \approx 1.4$ for HEWL [21]. More recently, an equation of state relating the spectral dimension, fractal dimension and the size of a protein was developed based on the coexistence of stability and flexibility in proteins [22].

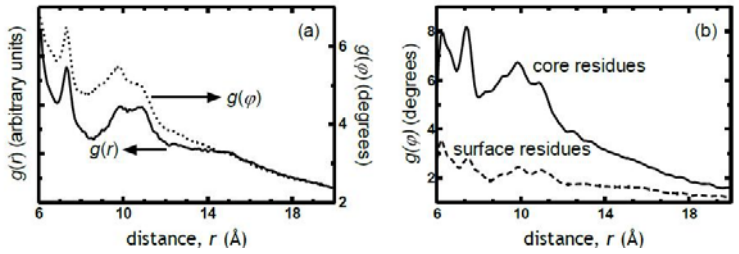

Figure 2. (a) Radial and angular distribution functions obtained by averaging over 595 proteins. (b) ADFs computed separately for the core and surface residues for a subset of 60 proteins.

In Fig. 3a, we display the r_c dependence of normal mode spectra averaged over 26 proteins of size 150 ± 10 residues, enabling us to disregard the size effect in the calculations. The low-frequency band of the graph is responsible for large amplitude collective motions related to function, whereas the high-frequency band refers to small amplitude motions of individual residues. We find that at r_c = 7 Å (where neighbors are from the first coordination shell), the distribution is characterized by a direct drop in density with increasing frequency. The universal behavior of the slow vibrational modes of proteins is recovered at higher r_c values. Above the cut-off distances that include the fourth coordination shell (r_c > 12 Å), a shoulder in the higher frequency region first appears, then broadens as r_c is increased. At r_c > 16 Å, a two-peaked density profile that is uncharacteristic of proteins sets in (inset to Fig 3a).

Thus, an r_c value in the range of 8 – 16 Å captures the general shape of protein vibrational spectra. Yet, inasmuch as one utilizes network models to study collective motions of proteins as a superposition of several low frequency modes, it is important to capture the distribution in the slow mode region of the protein in more detail. This region is intimately related to material properties, characterized by the spectral dimension, d_s. In Fig. 3b, we plot the spectral dimensions of these systems, obtained from power law best-fits to the cumulative density of modes, $G(\omega) \propto \omega^{d_s}$ for the first 70 modes in each set of data [with $dG(\omega)/d\omega = g(\omega)$]. The dimensions approach the Debye model value of 3 as r_c is increased (dotted line in the figure). The spectral dimensions in the r_c range from the second to the fourth coordination shell, (8 – 12 Å increase from below d_s = 1 to ca. d_s = 1.5. Furthermore, a crossover in the rate of change of the spectral dimension with the cut-off distance occurs at r_c = 16 Å, the slope reducing from ca. 0.13 to half this value; the crossover is accompanied by the shift to d_s > 2. Thus, it is plausible to use the cut-off value up to 16 Å so as to capture both the general shape of the vibrational spectra of proteins, as well as the spectral dimension that describes the density of slow modes.

Biological relevance. In recent years, network models of proteins, RNA and their complexes have opened up previously unprecedented areas of study, since the level of coarse graining adopted has been shown to describe several important phenomena unique to these self-assembled systems. The findings are mainly based on the observation that a simplified harmonic potential is

capable of describing the collective modes of motion [6], which also are associated with the basic functioning of these molecular machines [11]. The level of success of these studies in relation to the method of network construction has not been addressed systematically. We find for a number of proteins that the correlation between the mean-square fluctuations of C_α atoms and the theoretical predictions improve as the cut-off distance is increased. This curious observation is valid up to very large r_c values; i.e. for some proteins, even when all residues are interconnected, the fluctuations of individual residues are faithfully predicted. One example is displayed in Fig. 4 for a 263 residue β-class protein (PDB code: 1arb), where the residue-by-residue experimental B-factors (middle curve in gray in Fig. 4a) are compared with several selected theoretical models: A relatively low correlation is obtained at r_c = 8 Å; in particular, the fluctuations of surface loop residues 15 – 20 and 135 – 145 are overestimated due to the absence of important core-region contacts that are not taken into account at this r_c. The r_c = 15 Å model captures the experimentally determined fluctuation patterns, which remains unaltered at higher r_c. The Pearson correlation coefficients at a wide range of cut-off distances are plotted in Fig. 4b. We emphasize that the behavior exemplified by Fig. 4 is not unique to this protein, but is rather a common property of all proteins.

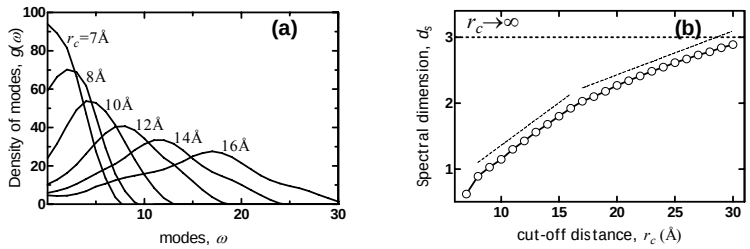

Figure 3. (a) Change of the density of vibrational modes, g(ω), with r_c. Inset displays the results for r_c values up to 30 Å. (b) Spectral dimension, d_s, of the networks, obtained from power law best-fits to cumulative density of modes. Goodness of fit is 0.98 or better in all cases. Thin dashed lines are included to guide the eye for the cross-over in the rate of change of d_s with r_c. Theoretical limit at d_s = 3 when all nodes are interconnected ($r_c \to \infty$) is also marked.

The increase in the correlation coefficient with r_c as well as its persistence to very high r_c values implies that the main ingredients that contribute to the fluctuation predictions are present in the Hessian obtained at a relatively low r_c, and the additional contacts act as a perturbation to this "essential" part of the matrix. We may thus partition the Hessian into two, where $\mathbf{H}_*$ contains information due to the essential contacts of the matrix whereas $\mathbf{H}_r$ is the residual part where the interactions are added in a spherically symmetrical manner around the nodes beyond a certain r_c value (Fig. 2). The inverse of the Hessian will be nominally modified, so that the predicted C_α fluctuations will change only slightly. A proof of this effect on the slow modes and the corresponding eigenvectors of the Hessian as well as the fits to detailed molecular dynamics simulations will be separately reported in a forthcoming manuscript.

Due to the invariance of the eigenvectors under a perturbation $\mathbf{H}_r$ to $\mathbf{H}_*$, the mode based predictions on the direction of motion between the unbound and bound conformations of the protein will also converge. An example is shown in Fig. 4c for the protein adenylate kinase, for which the eigenvector belonging to the slowest eigenvalue is known to describe the conformational change with high accuracy due to the highly collective behavior of the hinge motion between the two domains [23]. The Pearson correlation between the experimental and theoretical curves is 0.9 at r_c > 8 Å. The largest discrepancy between theory and experiment is

observed in residues 30 – 67 which belongs to the NMP binding domain closing over the ATP binding domain (called the LID) on the opposite side, the latter spanning residues 118 – 167. The prediction does not change with r_c.

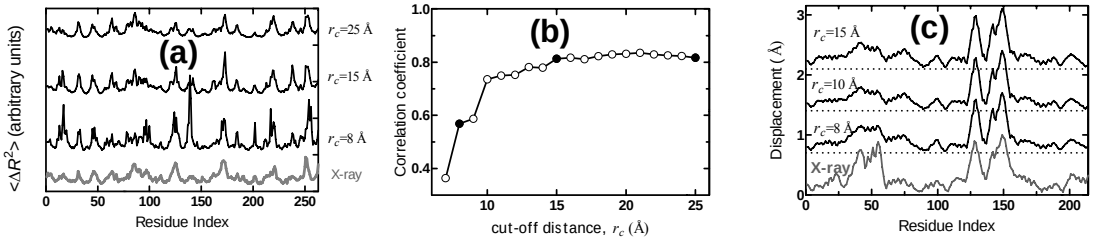

Figure 4. (a) Comparison of X-ray B-factors with predicted fluctuation profiles (PDB code: 1arb). (b) Pearson correlation coefficients at a wide range of cut-off distances for the protein in (a). (c) The displacement profiles of adenylate kinase in unbound and bound forms (PDB codes: 4ake and 1ake).

ACKNOWLEDGEMENTS

Financial support from the I2CAM (Institute for Complex Adaptive Matter) for the travel expenses to the MRS Fall Meeting is gratefully acknowledged (NSF Grant # DMR-0844115).

REFERENCES

1. A. Soyer, J. Chomilier, J. P. Mornon, R. Jullien, and J. F. Sadoc, Phys. Rev. Lett. **85**, 3532 (2000).
2. G. Raghunathan and R. L. Jernigan, Prot. Sci. **6**, 2072 (1997).
3. A. R. Atilgan, P. Akan, and C. Baysal, Biophys. J. **86**, 85 (2004).
4. J. Liang and K. A. Dill, Biophys. J. **81**, 751 (2001).
5. D. ben-Avraham, Phys. Rev. B **47**, 14559 (1993).
6. I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, Phys. Rev. Lett. **80**, 2733 (1998).
7. I. Bahar and A. J. Rader, Curr. Opin. Struct. Biol. **15**, 586 (2005).
8. Q. Cui, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. (Chapman & Hall/CRC, FL, USA, 2006).
9. I. Bahar, A. R. Atilgan, and B. Erman, Folding & Design **2**, 173 (1997).
10. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).
11. K. Hinsen, Proteins **33**, 417 (1998).
12. L. S. Yilmaz and A. R. Atilgan, J. Chem. Phys. **113**, 4454 (2000).
13. A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).
14. P. Doruker, A. R. Atilgan, and I. Bahar, Proteins **40**, 512 (2000).
15. L. Yang, G. Song, and R. L. Jernigan, Biophys. J. **93**, 920 (2007).
16. P. Petrone and V. S. Pande, Biophys. J. **90**, 1583 (2006).
17. P. Fariselli and R. Casadio, Prot. Eng. **12**, 15 (1999).
18. H. M. Berman et al., Nucl. Acids Res. **28**, 235 (2000).
19. Y. Q. Zhou, D. Vitkup, and M. Karplus, J. Mol. Biol. **285**, 1371 (1999).
20. Charles Kittel, *Introduction to Solid State Physics*. (John Wiley & Sons, 2004), 8th ed.
21. A. V. Svanidze et al., Ferroelectrics **348**, 556 (2007).
22. S. Reuveni, R. Granek, and J. Klafter, Phys. Rev. Lett. **100**, 4 (2008).
23. F. Tama, W. Wriggers, and C. L. Brooks, J. Mol. Biol. **321**, 297 (2002).