Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 22, Issue Number 4, (2005)

Protein Interactions and Fluctuations in a Proteomic Network using an Elastic Network Model

http://www.jbsdonline.com

Melik C. Demirel[1,*]
Ozlem Keskin[2,3]

[1]College of Engineering
Pennsylvania State University
University Park, PA 16802 USA
[2]Center of Computational Biology and Bioinformatics and College of Engineering
Koc University
34450, Istanbul, Turkey
[3]Laboratory of Experimental and Computational Biology
NCI-Frederick, NIH
Frederick, MD 21702 USA

Abstract

A set of protein conformations are analyzed by normal mode analysis. An elastic network model is used to obtain fluctuation and cooperativity of residues with low amplitude fluctuations across different species. Slow modes that are associated with the function of proteins have common features among different protein structures. We show that the degree of flexibility of the protein is important for proteins to interact with other proteins and as the species gets more complex its proteins become more flexible. In the complex organism, higher cooperativity arises due to protein structure and connectivity.

Introduction

Proteins act alone or in complexes to perform many cellular functions. Protein-protein interactions are rather complex. There are approximately 4200 genes in *yeast* that end up approximately 93,000 protein-protein interactions (1, 2). Another level of complexity arises when we start thinking about protein complexes, *e.g.*, dimers, trimers, and tetramers. However, the number of proteins in a complex is not limited to such lower numbers, because there are complexes that have more than 50 different proteins. For example, eukaryotic ribosomes contain 82 integral proteins, and yeast proteasomes plus their ancillary components comprise 56 polypeptides. Three hundred and twenty-six proteins were recently identified being associated with the RNA polymerase II preinitiation complex from *Saccharomyces cerevisiae* (3).

The number and type of interactions in proteins are very complex, and protein 3 dimensional structure plays a key role in these interactions. Mutations and adaptations provide the proteins through evolution to have marginal stability. Besides, some residues and regions are conserved in both protein sequence and structure. Details of the protein topology in the folded state have been studied by several groups (see for a review: Taylor *et al*. (4)). Freire and coworkers (5) studied the flexibility of substrates that bind to HIV protease. They observed that peptide substrates are more flexible than the synthetic peptides and they could adapt themselves better to bind to the HIV protease. From computational point of view, one frequently used method involves molecular dynamics (MD) simulations following the motion and folding of proteins at the molecular level. While these techniques are very effective, their high computational cost is a drawback. A viable alternative method is the coarse-grained simulations for describing the vibrational dynamics of simple models (6-10).

Fernandez and colleagues (11) showed that the number of dehydrons in a protein increases as the species gets more complex. In this paper, we perform a normal mode analysis to determine the vibrational motions of proteins across different species. We studied the fluctuation of residues with low amplitude fluctuations in slow modes. We observe that in all cases, a protein that belongs to complex organism has more correlated motion in its vibrational motions than the protein in the simpler organism.

*Phone: 814-863-2270
E-mail: melikdemirel@psu.edu

Materials and Methods

Our method, which is now known as the Gaussian Network Model (GNM), models fluctuations of proteins. The GNM method is very successful in describing the dynamic characteristics of proteins (12-18). Comparison with experiments shows that slow and fast modes of proteins are associated, respectively, with function and stability (12). Results from GNM calculations were found to be in excellent agreement with x-ray crystallographic temperature factors (also called Debye-Waller or B-factors) (12, 19). The GNM is based on the following postulate: in folded proteins, residues undergo gaussianly distributed fluctuations around their mean positions, due to harmonic potentials between all "contacting" residues. No residue specificity need be invoked as a first order approximation. Instead, the inter-residue potentials are all represented by the same single-parameter harmonic potential. The fluctuations of residues are controlled by a harmonic potential and α-carbons being used as representative sites for residues. The dynamic characteristics of the molecule are fully described in this model by the so-called Kirchhoff matrix of contacts. Two residues are defined to be in contact if the distance between their α-carbons is less than a cut off radius of 8 Å. Kirchhoff matrix of contacts and harmonic potential are defined as

$$G = \begin{cases} -\delta(r_c - r_{ij}) & i \neq j \\ -\sum\Gamma_{ij} & i = j \end{cases} \qquad H = {}^1\!/_2\ \gamma(\Delta R^T \Gamma \Delta R) \qquad [1]$$

where ΔR is the fluctuation of an α-carbon atom and Γ is the Kirchhoff matrix (or the contact map). Note that the generalized inverse of the Kirchhoff matrix is taken here after eliminating the zero eigenvalue. Fluctuations of residues are obtained by inverting the Kirchhoff matrix and given by

$$\langle \Delta R_i \Delta R_j \rangle = {}^3\!/_\gamma\ k_B T\, [\Gamma]_{ij}^{-1} \qquad [2]$$

where k_B is the Boltzmann constant and T is the absolute temperature. $\langle \Delta R \Delta R^T \rangle$ can be expressed as a sum over the contributions $\langle \Delta R_i \Delta R_j^T \rangle_k$ of the individual modes, in an expansion using the eigenvalues λ_k and eigenvectors U_k of Γ in

$$\langle \Delta R_i \Delta R_j \rangle = \sum_k \left[\Delta R_i \Delta R_j \right]_k = \left({}^3\!/_\gamma\ k_B T \right) \sum_k \left[\lambda_k^{-1} u_k u_k^T \right]_{ij} \qquad [3]$$

Here, the summation is performed over all ($0 < k < N$) nonzero eigenvalues of Γ. A first test of the validity of the GNM is to compare the predicted fluctuations of residues with those observed in experiments (B-factors). We extracted a set of 3-D structures for all single chain protein types from protein data bank, and calculated free energies of 2656 proteins. Later, we selected a non-redundant subset of 302 proteins with at most 50% sequence similarity. A force constant of 1.0 ± 0.5 kcal/(mol Å^2) has been obtained in 302 proteins which agrees with unpublished results of Bahar *et al*. (20) In protein crystals, temperature factors are of the order of 12-20Å^2, corresponding to a displacement of the atoms about their mean positions of between 0.15 and 0.5Å (21). However, in almost all crystal structures, temperature factor is determined empirically and also takes into account a variety of other factors such as static disorder, wrong scaling of measurements, absorption, and incorrect atomic scattering curves. Hence values of temperature factors and force constant may only be taken as a rough approximation.

Results and Discussion

We have studied a number of small size proteins from protein data bank, PDB, (Chymotrypsin Inhibitor 2 pdb.2ci2, Muscarinic Toxin/Acetylcholine Receptor

Binding Protein pdb.1ff4, Type III Antifreeze Protein,pdb.1gzi) which have no similarity in their protein structures but have the same number of residues (N=65). They are all observed to exhibit similar eigenvalue distributions by using GNM. We have also studied the fluctuation distributions for these proteins using slow mode analyzes of GNM (Figure 1a).

Fluctuations are represented by a color code that goes from red to blue, where blue is the most flexible and the red is the most rigid. These three proteins exhibit very different fluctuations but their eigenvalue distributions are similar (Figure 1b). Based on this observation, we studied a set of 3-D structures from 302 proteins and calculated the connectivities of these proteins. GNM model defines the protein connectivity by the following equation,

$$connectivity \equiv \sum_i \Gamma_{ii} = \sum_k \lambda_k \qquad [4]$$

where Γ is the contact map, and λ is the eigenvalue which are introduced in the previous section. The results from 302 proteins showed that the connectivity of protein structures scales linearly with protein length (Figure 2). Combining the results of Eq. [4] and Figure 2, we can conclude that the proteins which have equal number of residues have the same connectivity value and the sum of eigenvalues is also same. Similar eigenvalue distribution is expected for proteins which have equal number of residues, but the fluctuations will be different based on topological constrains. This should be an outcome of the fact that a global protein has a characteristic packing in its cores. On the average, a residue will have seven non-bounded neighbors in its first coordination shell (22). Therefore, slow modes should have common features among different protein structures.

The interconnection between the folds and functions is a well known fact, such that folds across different species are conserved. It is also known that a complex and a simple organism have almost similar number of genes. The complexity of the organisms should be the result of the interactions between these genes' products, namely their proteins. Although the folds are similar, not all structural features are conserved across different species. Fernandez and colleagues (11) showed that the number of dehydrons in a protein increases as the species gets more complex. These sites are assumed to be interaction sites with other proteins. This also suggests that there should be differences among the dynamics of proteins across species.

We analyzed all the data related to the fluctuations for the same folding domains across different species (Table I). The first column gives the names of the proteins. The second column lists the two PDB codes for the same protein; the first one belongs to a complex organism and the second one to a simpler organism. Third and Fourth columns are the residue numbers in the proteins taken from the PDB. ρ_1 and ρ_2 refer to the ratio of the residues that exhibit high fluctuations to the overall residue numbers for the higher organism and simpler organism, respec-

Table I
Ratios of highly flexible residues for the same species for different complexities.

Protein	PDB code	N_1	N_2	ρ_1*	ρ_2*	F_1#	F_2#
Prion	1qm0,1ag2	104	103	0.26	0.19	0.36	0.30
Ubiquitin	1ubi, 1f0z	76	66	0.33	0.14	0.26	0.23
SH3 protein	5hck,1sem	61	58	0.23	0.17	0.22	0.20
Hemoglobin	1gob,1dlw	155	116	0.18	0.12	0.26	0.24
Chaperone	1byq,1amw	213	213	0.19	0.16	0.68	0.63
Myoglobin	2hbc,1bz6	141	153	0.37	0.33	0.19	0.19

*ρ_1 and ρ_2 refer to the ratio of the residues that exhibit high fluctuations to the overall residue numbers for the higher organism and simpler organism, respectively.
#F_1 and F_2 are the average residue fluctuations over all modes for the higher organism and simpler organism, respectively. N_1 and N_2 are the residue numbers in the organisms.

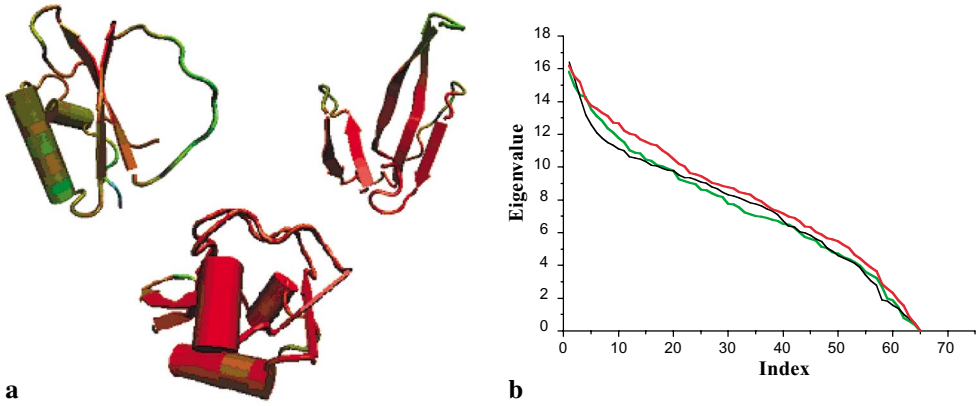

Figure 1: GNM fluctuation analysis of small proteins (PDB codes: 2ci2, 1ff4, 1gzi) which has the same length but different structure are shown in (**a**). Fluctuations are represented by a color code that goes from red to blue, where blue is the most flexible and the red is more rigid. Eigenvalue distributions for these proteins are shown in (**b**).

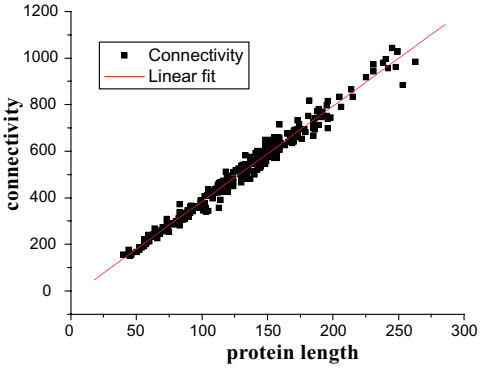

Figure 2: Connectivity value from 302 x-ray structures which are extracted from PDB. Connectivity is linearly varying with the protein length.

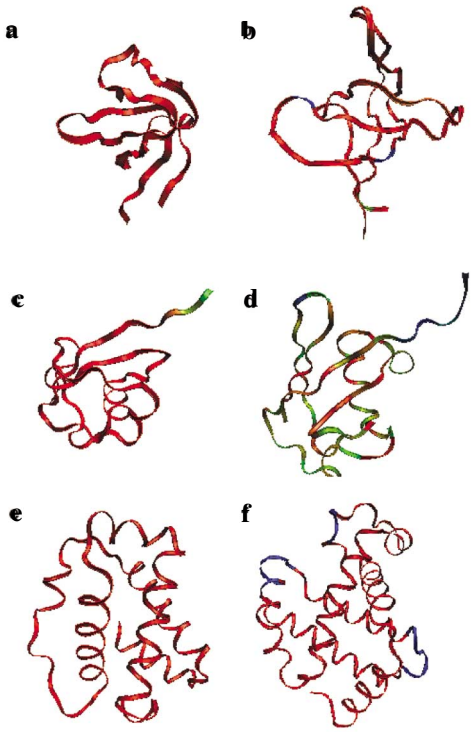

Figure 3: Illustrative comparative analysis of the packing flexibilities for the same proteins in different species. The highly flexible regions are shown in blue and the least flexible with red. The left and right panels represent the less and more complex proteins, respectively. SH3 domains are from nematode *C. elegans* (pdb.3sem) (**a**) and *H. sapiens* (pdb.5hck) (**b**) ; ubiquitin is from *E. coli* (pdb.1f0z) (**c**) and *H. sapiens* (pdb.1ubi) (**d**); and hemoglobin is from *Paramecium* (pdb.1dlw) (**e**) and *H. sapiens* subunit (pdb.1bz0, chain B) (**f**).

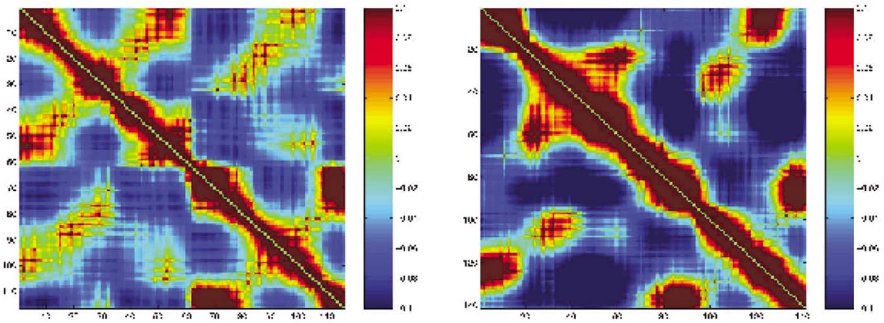

Figure 4: The cooperativity maps of hemoglobin for *Paramecium* (**a**) and *H.sapiens* (**b**) are shown.

tively. A residue is flagged as a highly fluctuating residue if its average fluctuation over the first three slow modes is above its average fluctuation. The average fluctuation is calculated from Eq. [3], by taking $1 \leq k \leq 3$. Therefore, these numbers are simply indicators of whether a protein has a high number of flexible residues or not. The seventh and eighth columns are the average residue fluctuations over all modes for the higher organism and simpler organism, respectively. The average fluctuations are calculated from Eq. [3], by taking $0 < k < N$. In all cases, we can see that the more complex the organism the more flexible its proteins. The more complex organisms also have higher ρ values.

We observe that when the same protein within a protein family is examined across different species, there are differences in their flexibilities. For example, the Src homology 3 (SH3) proteins in the nematode *Caenorhabditis elegans* (pdb.3sem) has less flexibility compared to the human SH3 domain (pdb.5hck) (Fig. 3a). Since flexibility determine protein recognition sites and its ability to bind to other molecules, this difference suggests a far more complex recognition mechanism in the complex species. Likewise, on the same basis, the human ubiquitin (pdb.1ubi) is more flexible than its *Escherichia coli* counterpart (pdb.1f0z) (Fig. 3b). For the hemoglobin protein, there are significant differences: the *paramecium* "hemoglobin" (pdb.1dlw) is more rigid and is monomeric *in vivo*. The analogous-fold hemoglobin (pdb.1bz0) in humans exhibits higher degree of flexibility (Fig. 3c) and occurs as a tetramer. Figure 3 and Table I suggest that as more complex species diverge, the conserved fold associated with a given function becomes more interactive as suggested by Fernandez and colleagues (11). Through out the evolution we observe increasing number of binding-related residues.

Figure 4 shows the cooperativity maps of two hemoglobin proteins from *Paramecium* and *H.sapiens* that are calculated from Eq. 2. Cooperativity map are plotted by a color code that goes from red to blue, where red is the positively correlated residues, and the blue is negative. In addition, yellow color denotes no correlation among residues. Higher complexity organism (*H.sapiens*) has more red and blue regions than the lower complexity one. In the complex organism, higher cooperativity arises due to protein structure and connectivity.

The same protein fold is mostly utilized for function across different species. On the other hand, the degree of cooperativity of a protein with other molecules is much higher in complex organisms making it to form more complex pathways. Therefore genotypic differentiation brings about variability in the extent of molecular association. This variability arises from differences in the extent to which intramolecular hydrogen bonds are shielded from water attack as suggested by Fernandez and colleagues (11). Here we also show that, the degree of flexibility of the protein is also important for proteins to interact with other proteins and as the species gets more complex its proteins become more flexible. Higher levels of connectivity are advantageous for certain functions in more complex species. The higher organisms remain their complex physiologies without dramatically increasing their genome size (the number of genes in the human genome proved to be deceptively low). Our results imply that the interactions in a species are determined by the flexibility in its protein folds and this might be an indication of the complexity, helping explain how complex physiologies may be achieved without a significant increase in genome size.

Acknowledgement

This research is supported by MRSEC/MRI seed grant of Pennsylvania State University funds (to M.C.D.), Institute for Complex Adaptive Matter fellowship (to M.C.D.), and National Institutes of Heath intramural funds (to O.K.). We thank Jayanth Banavar at the Pennsylvania State University for discussions.

References and Footnotes

1. J. E. Galagan, S. E. Calvo, K. A. Borkovich, E. U.Selker, N. D. Read, D. Jaffe, W. FitzHugh, L. J. Ma, S. Smirnov, S. Purcell *et al*. *Nature 422*, 859-868 (2003).
2. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart *et al*. *Nature 403*, 623-627 (2000).
3. J. Ranish, E. Yi, D. Leslie, S. Purvine, D. Goodlett, J. Eng, and R. Aebersold. *Nat. Genet. 33*, 349-355 (2003).
4. W. R. Taylor, A. C. W. May, N. P. Brown, and A. Aszodi. *Reports on Progress in Physics 64*, 517-590 (2001).
5. I. Luque, M. J. Todd, J. Gomez, N. Semo, and E. Freire. *Biochemistry 37*, 5791-5797 (1998).
6. I. Bahar, A. Atilgan, M. C. Demirel, and B. Erman. *Physical Review Letters 80*, 2733-2736 (1998).
7. F. Tama. *Protein and Peptide Letters 10*, 119-132 (2003).

8. R. Lumry. *Biophysical Chemistry 101*, 81-92 (2002).
9. O. Keskin, S. R. Durell, I. Bahar, R. L. Jernigan, and D. G. Covell. *Biophysical Journal 83*, 663-680 (2002).
10. C. Micheletti, J. R. Banavar, and A. Maritan. *Physical Review Letters 8708* (2001). Art. no.-088102.
11. A. Fernandez, R. Scott, and R. S. Berry. *Proceedings of the National Academy of Sciences of the United States of America 101*, 2823-2827 (2004).
12. M. C. Demirel, A. Atilgan, R. L. Jernigan, B. Erman, and I. Bahar. *Protein Science 7*, 2522-2532 (1998).
13. R. L. Jernigan, M. C. Demirel, and I. Bahar. *International Journal Of Quantum Chemistry 75*, 301-312 (1999).
14. A. Atilgan, S. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. *Biophysical Journal 80*, 505-515 (2001).
15. I. Bahar, A. D. C. Wallquist, and R. Jemigan. *Biochemistry 37*, 1067-1075 (1998).
16. O. Keskin, I. Bahar, D. Flatow, D. Covell, and R. Jernigan. *Biochemistry 41*, 491-501 (2002).
17. O. Keskin, X. Ji, J. Blaszcyk, and D. Covell. *Proteins 49*, 191-205 (2002).
18. O. Keskin. *J Biomol Struct Dyn 20*, 333-345 (2002).
19. M. C. Demirel, I. Bahar, and A. Atilgan. *Biophysical Journal 76*, A176 (1999).
20. I. Bahar and B. Ozkan. Unpublished Data. (2002). http://klee.bme.boun.edu.tr/gamma.
21. X. Ji. National Institutes of Health, Personal Communication (2003).
22. S. Miyazawa and R. L. Jernigan. *Journal of Molecular Biology 256*, 623-644 (1996).

Date Received: August 18, 2004

Communicated by the Editor Ramaswamy H Sarma