Available online at www.sciencedirect.com

SCIENCE DIRECT®

Nanomedicine: Nanotechnology, Biology, and Medicine 1 (2005) 41–46

nanomedicine

www.nanomedjournal.com

Basic Research

Clustering and diversity of fluctuations for proteins

Melik C. Demirel, PhD,[a,b,*] Dmitry Cherny, PhD[b]

[a]College of Engineering, Pennsylvania State University, University Park, Pennsylvania

[b]Department of Molecular Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Received 16 October 2004; accepted 25 November 2004

Abstract

Background: Protein topology plays a key role in various types of interactions. Topological constraints of a protein are defined by a contact map. We studied the fluctuations of proteins with use of a new approach based on contact map.

Methods: An annealing algorithm is used to generate a 3-dimensional protein structure from the contact map. First, we study the properties of structural elements based on fluctuations by adding individual structures (domains or subdomains). Thereafter, we focus on the building block of proteins in terms of fluctuations.

Results: To verify our hypothesis, we analyzed the pattern of fluctuations for chymotrypsin inhibitor-2 (CI2) by unstructuring (melting) of subregions. The data show different patterns of fluctuations for the unstructured CI2 relative to that calculated for the intact protein.

Conclusion: Our approach introduces a new concept for classifying building blocks of proteins based on thermal fluctuations.

Key words: Protein fluctuations; Elastic network model; Simulated annealing

Proteins exhibit various types of interactions, and protein topology plays a key role in these interactions. Details of the protein topology in the folded state have been studied by several groups (for a review, see Taylor et al [1]). Topologic constraints of a protein are defined by a contact map. Several simulated annealing algorithms—for example, XPLOR-NIH [2] and Rosetta [3]—are used together with nuclear magnetic resonance (NMR) experiments for obtaining protein structure(s) based on distance constraints. Vendruscolo et al [4] have introduced an annealing algorithm to generate a 3-dimensional protein structure from the contact map. Hu et al [5] introduced a data mining algorithm to characterize contact maps for different proteins, and to search for patterns which may be used for the contact map prediction of unknown proteins. Park and Levitt [6] have created protein conformations that possess some characteristics of native proteins (so-called decoy proteins).

Particular folding pathways may favor distinct types of topology [7], and a topologic similarity between structures might imply an evolutionary relationship [8]. Efimov [9] has considered how protein folding patterns may be built up from basic supersecondary motifs. Typical folding patterns were classified by Orengo et al [10], and they have observed that ~80% of known protein structures fall approximately into ~20% of distinct folding patterns. Taylor [11] opened up a new approach to the classification of protein structures by introducing a set of idealized structures that are compared with all known structures.

In this article, based on an elastic network model [12] and a simulated annealing algorithm [2], we analyzed the clustering and diversity of protein fluctuations on the basis of two views. First, we studied the fluctuation properties of secondary structural elements. After this, we focus on the building block of proteins in terms of fluctuations.

Material and methods

We have developed a computational tool for understanding the equilibrium fluctuations of proteins. Our method, which is now known as the gaussian network model (GNM), models the fluctuations of proteins displaying an excellent agreement with x-ray crystallographic temperature factors (also called Debye–Waller or B-factors) [12,13]. The

[*] Corresponding author. 212 EES Bldg., School of Engineering, Pennsylvania State University, University Park, PA 16802.

E-mail address: melikdemirel@psu.edu (M.C. Demirel)

1549-9634/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.nano.2004.11.007

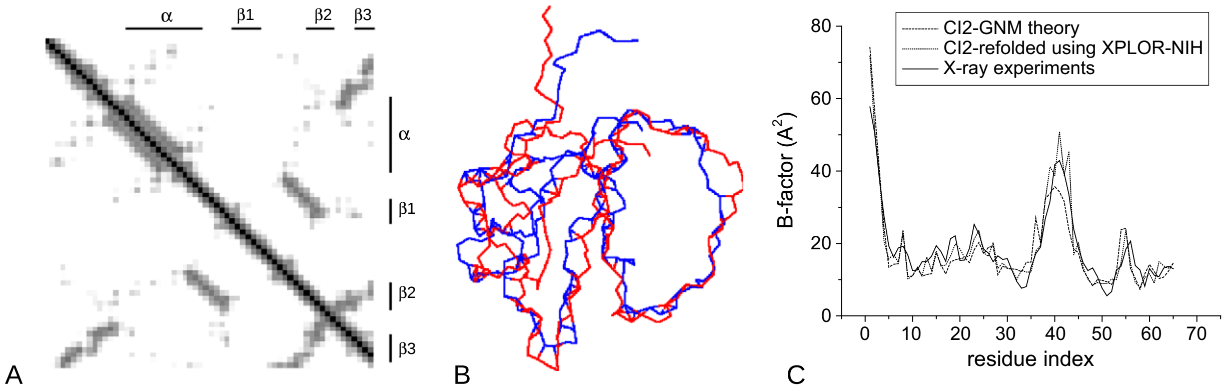

Fig 1. **A**, Contact map for the X-ray crystallographic structure of CI2. **B**, Computationally refolded (red) and x-ray structures of CI2 structure (blue). **C**, α-carbon fluctuations of CI2 as determined from x-ray data (B-factors, solid), gaussian network model approach (dotted), and computationally refolded structure (dashed).

GNM method is very successful in describing the dynamic characteristics of proteins [12-15]. Comparison with the experimental data shows that slow and fast modes of protein motion are associated with its function and stability, respectively [16,17].

The GNM method is based on the following assumption: in folded proteins, residues undergo fluctuations that exhibit gaussian distribution around the mean positions, due to harmonic potentials between all "contacting" residues (for a detailed explanation, see Demirel et al [12,13], Atilgan et al [14], and Bahar et al [15]). No residue specificity needs to be invoked with the first order of approximation. Instead, the interresidue potentials are all represented by the same single-parameter (γ) harmonic potential. The fluctuations of residues are controlled by a harmonic potential with α-carbons being used as representative sites for residues. The dynamic characteristics of the entire protein molecule are fully described by the so-called Kirchhoff matrix of contacts. Two residues are defined to be in contact if the distance between their α-carbons is lower than the cutoff radius (r_c) of 7 Å [12]. The Kirchhoff matrix of contacts and harmonic potential are defined as follows:

$$\Gamma = \left\{ \begin{array}{ll} -\delta(r_c - r_{ij}) & i \neq j \\ -\sum \Gamma_{ij} & i = j \end{array} \right\} \qquad (1)$$

$$H = {}^{1}\!/_{2}\gamma(\Delta R^T \Gamma \Delta R),$$

where ΔR is the fluctuation of an α-carbon atom and Γ is the Kirchhoff matrix or, the contact map. The abovementioned equations follow from the integration of the single parameter multivariate Gaussian function in the configurational integral, originally given by Flory [18]. Note that the generalized inverse of the Kirchhoff matrix is taken here after eliminating the zero eigenvalues. Fluctuations of residues are obtained by inverting the contact map and given by

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

where k_B is the Boltzmann constant and T is the absolute temperature. The GNM model defines the protein connectivity by the trace of Kirchhoff matrix,

$$connectivity \equiv \sum_i \Gamma_{ii}\,. \qquad (3)$$

The contact map of a protein can be obtained from the 3-dimensional structure using Eq 1. However, the inverse of this procedure (eg, generating 3-dimensional structure from the contact map) is not trivial. We have used an annealing algorithm, XPLOR-NIH, to generate a 3-dimensional protein structure from the contact map. XPLOR-NIH generates an ensemble of structures based on the topologic constraints (for a detailed explanation, see Brunger et al [19]). XPLOR's main focus is the 3-dimensional structure determination of macromolecules using crystallographic diffraction or NMR data. XPLOR-NIH was originally derived from XPLOR version 3.8 and contains all of the functions therein [19]. The program is based on an energy function approach: arbitrary combinations of empirical, geometric, and effective energy terms describing experimental data may be used. The combined energy function can be minimized by a variety of gradient descent, simulated annealing, and conformational search procedures. Optimizing the atomic coordinates to match the NMR observables can be achieved by several methods. In Cartesian coordinates, XPLOR-NIH provides Powell gradient minimization, and annealing optimization simulated by molecular dynamics [2].

Results

We have applied our approach to chymotrypsin inhibitor-2 (CI2), which has been extensively studied experimentally [20] and by means of computer modeling [21]. Structurally, the CI2 consists of 2 main domains; the first is an extended N-terminus linked to an α-helix (residues 12 to 24), and the second is made up of 3 β-sheets (residues 28 to 34, 45 to 52, and 60 to 64, respectively) and the reactive site loop (residues 35 to 44). In this work, the

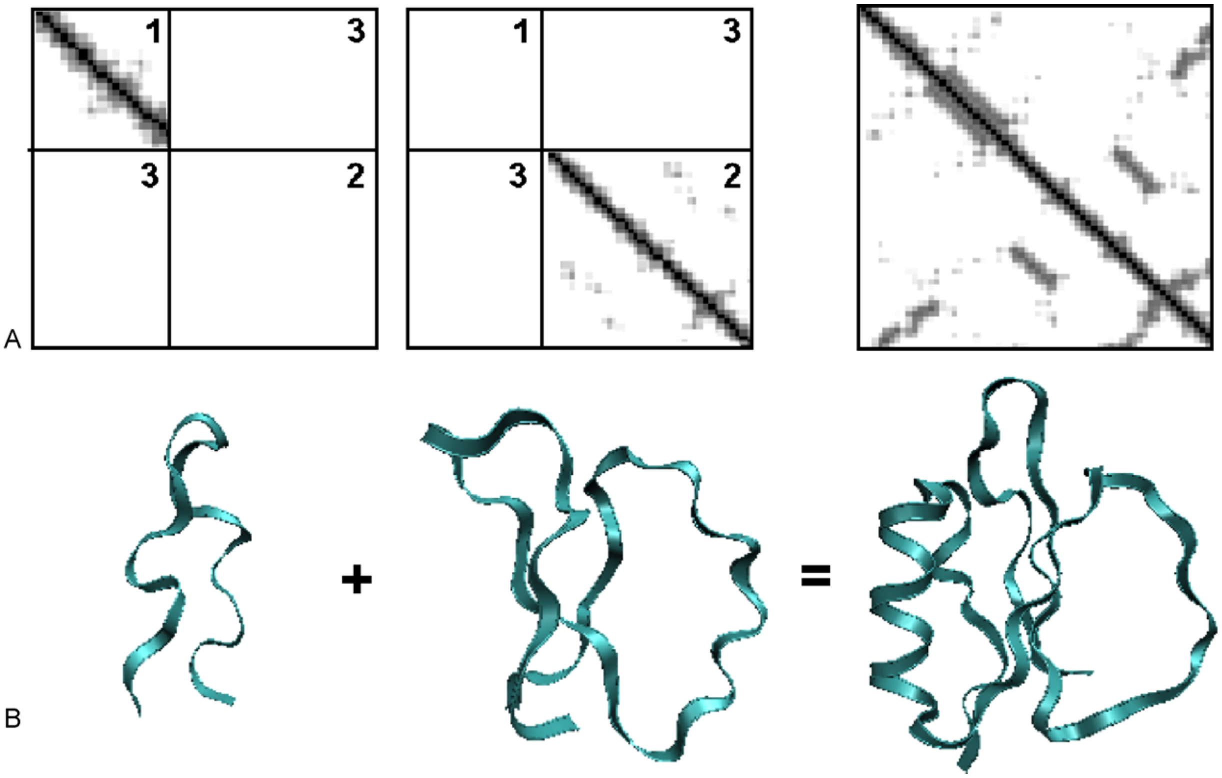

Fig 2. Contact maps (**A**) and corresponding refolded structures (**B**) are shown for region 1 (residues 1 to 24), region 2 (residues 25 to 64), and for the entire CI2 protein (residues 1 to 64). Note that the structures for both regions can be simply combined, thus forming the entire structure.

original residues of CI2 (20 to 83) are renumbered (1 to 64). The hydrophobic core of the protein is formed by a delicate association of the β-sheet and the α-helix. It consists of 10 hydrophobic residues: W5, L8, A16, I20, I29, V47, L49, V51, I57, and P61. It was found that major parts of hydrophobic residues are crucial for the formation and/or stability of the tertiary fold [12]. Furthermore, these residues participate in the formation and stabilization of a folding nucleus, as CI2 follows a 2-state folding kinetics model [22].

The simulated annealing technique is applied to the CI2 protein, and an ensemble of folded structures is computed using the XPLOR-NIH program. Folded structures exhibit a high degree of similarity to the original protein structure obtained from the protein data bank (PDB). We measure the similarity by the sum of root mean square deviations (RMSDs) for all α-carbons between the experimental and the simulated structures. Using a cutoff radius, r_c, of 7 Å (see Eq 1), the mean value of RMSD over an ensemble of 1000 simulated structures was found to be 3.20 Å (SD, 0.62 Å), whereas the minimum and maximum values were 2.24 Å and 5.02 Å, respectively. Increasing the cutoff radius up to 9 Å allowed getting a structure with lower RMSD equal to 1.60 Å.

The contact map for C12 is shown in Figure 1, *A*. The secondary structure elements are shown along the axes of Figure 1, *A*. We generate the 3-dimensional protein structures from the contact map using the simulated annealing algorithm (XPLOR-NIH). The structure exhibited the lowest RMSD (red) relative to the PDB structure (blue) are presented in Figure 1, *B*. In addition, temperature factors (B-factors) for CI2 are displayed in Figure 1, *C*. Calculated values using the GNM approach and the values of B-factors determined from x-ray data exhibit a very good agreement. The most prominent peak in Figure 1, *C*, is located at the position 35 to 44 (residue index) and does correspond to the binding region of the CI2. It is worth noting that refolded structure with the lowest value of RMSD, calculated from the contact map, also exhibit a very similar pattern of α-carbon fluctuations along the residue index (Figure 1, *C*).

The stepwise addition of secondary structures was introduced by Efimov [9] (and references therein). Common structural motifs in proteins are of particular value in protein modeling and design because they can be taken as the initial starting structures. Furthermore, stepwise addition of α-helices and/or β-strands to an initial structural motif can provide insights about the fold diversity and evolution. This is a mechanistic point of view in the sense that structures (eg, domains or subdomains) can be added to generate a final structure. For example, CI2 has 2 subdomains, where one is an extended N-terminus linked to an α-helix, and the other is made up of the β-sheets and the reactive site loop (region 1 and region 2, respectively, in Figure 2). We have separated these 2 subdomains and refolded each subdomain using the annealing algorithm (Figure 2, *A* and *B*). Region 3 is the interface between regions 1 and 2 and is required for

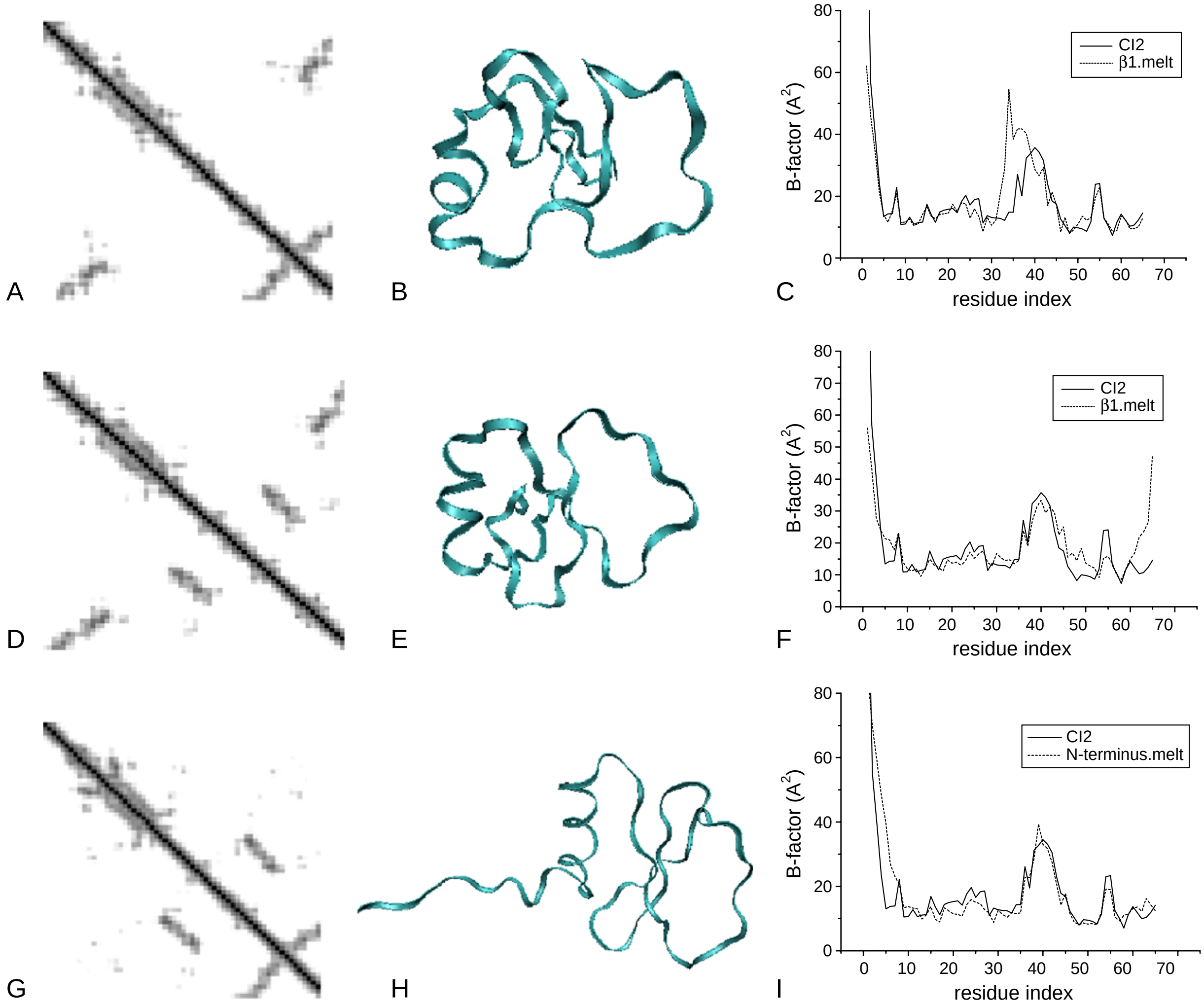

Fig 3. β1 region (named β1.melt, residues 28 to 34), β2 region (named β2.melt, residues 45 to 52) and N-terminus (named N-terminus.melt, residues 1 to 12) are shown as contact maps (**A**, **D**, and **G**, respectively), and as 3-dimensional structures (**B**, **E**, and **H**, respectively) using ribbon view of VMD visualization software [25]. Fluctuations for the entire protein along the residue index calculated for the intact structure using gaussian network model (solid), and for that after melting of β1, β2, and the N-terminus regions are presented in **C**, **F**, and **I** (dashed), respectively.

the formation of additional secondary structures and the stability of the entire structure (eg, the hydrophobic core). These two independent subdomains, which generate complementary parts, are added to build up the complete CI2 protein. The results demonstrate that the stepwise addition of secondary structures makes it possible to construct diverse structural elements (eg, subdomains, domains, or common folds).

Fluctuation properties of structural elements (eg, secondary structures) can be analyzed from a thermodynamic point of view. It is known that α-carbon fluctuations, which in general reflect their thermal motion, exhibit a nonuniform behavior relative to the α-carbons index (a pattern of fluctuations) [14]. This is due to the protein topology and/or interactions between amino acid residues. In general, fluctuations of hydrophobic core regions are lower than the surface residues since internal residues have more topologic constraints with respect to their surface-exposed counterparts. An example of this pattern for the CI2 protein is presented in Figure 1, *C*, showing regions of low and high fluctuations, which apparently can be grouped (along the residue index) depending on the level of fluctuations in each group. We introduce a new idea concerning the fluctuation properties of structural elements. Our conjecture is based on the hypothesis that the structural units constituting the entire protein differ in their fluctuations, at least at the level of α-carbons. To verify our hypothesis, we analyzed the pattern of fluctuations for the entire CI2 protein by unstructuring (melting) of α, β1, β2, and N-terminus regions, corresponding to the residues 12 to 24, 28 to 34, 45 to 52, and 1 to 12, respectively. These regions or folds can be individually unstructured (or melted) and subsequently refolded, giving rise to both the original structure and the level of fluctuations. To achieve this, we removed the contacts of the secondary structure by altering the contact map of CI2 that corresponds to each selected region and refolded the entire protein using simulated annealing. Thus, we obtained structures that differ from the intact

Table 1
Statistical values for computational folded minimum rmsd structures

Structure name	Connectivity at the hydropholic core*	Energy/residue (MJ-potential/64)	Connectivity	rmsd†
β1 melt	7.6	−7.73	221	0.48
β2 melt	7.2	−8.02	214	0.49
α-Helix melt	7.6	−11.82	215	0.40
N-terminus melt	8.0	−9.80	223	0.41
Intact protein	8.2	−14.88	233	0.37

* $\frac{1}{10}\sum_i \Gamma_{ii}$ where i is the 10 hydrophobic residues (5,8,16,20,29, 47,49,51,57,61) for the "minimum rmsd" structure.

† SD is calculated from all computationally generated structures.

structure as a result of the presence of regions lacking the α-carbon atoms contact (melted regions), and the corresponding fluctuations are depicted in Figure 3.

The data show 2 different patterns of fluctuations for refolded proteins—that is, containing melted regions—relative to that calculated for the intact protein. First, both patterns are very similar, as shown in Figure 3, *I*, implying that melted region does not influence the level of fluctuations of the adjacent regions. We should note that higher fluctuations at the N-terminal in Figure 3, *I*, are due to unstructured topology. Another situation is met for the β1 region. Its melting affected the level of fluctuations of the downstream residues, that is, residues located at 32 to 39, but not the upstream residues. Melting of the β2 region changes the fluctuation level of residues located at 44 to 58 and 60 to 64.

Global structural modifications of the proteins containing individual melted regions characterized by the protein connectivity (calculated for the hydrophobic core and entire molecule) and free energy (calculated by using Miyazawa-Jernigan potential [23]) are summarized in Table 1. Intact protein, corresponding to the crystal structure from PDB, has the lowest energy as expected and the highest values for the entire connectivity. However, depending on the importance of the secondary structures, the energy varies. For example, melting of the β1 sheet (β1 melt structure in Table 1) has a drastic effect on both the hydrophobic core and the overall energy of the protein. We have also calculated the refolding of α-helix structure, but we conserved the long-range interactions, which include 2 critical residues A16 and I20 at the hydrophobic core. The structure kept its stability and its energy value close to the intact protein, although the local connectivity at the hydrophobic core is the same as for the β1 melt structure.

Discussion

On the basis of our data of individually unstructured (or melted) folds to the fluctuations of the entire protein, we introduce a "building block" definition. If there is a difference in the pattern of fluctuations of the protein resulting from melting of an individual region, then the region can be defined as a protein building block in terms of fluctuations (eg, regions β1 and β2 fall into this category). If there is no difference in fluctuations, then the structural elements are not part of the building block, such as the N-terminus (Figure 3, *H* and *I*).

Fluctuations related to function govern the behavior of the protein, as in the processes of protein interaction (recognition) with other biologic molecules or ligands. For example, CI2 binds subtilisin novo through the recognition loop, and residue fluctuations (residues 35 to 44) are altered on binding [24]. Our approach allows for classification of building blocks of proteins based on their thermal fluctuations, which will eventually lead to a fluctuation database of proteins. We have applied our method to other sets of proteins (eg, hemoglobin, RMSD = 1.6 Å; ubiquitin, RMSD = 1.8 Å), and these results will be the main focus of future work.

Acknowledgments

We thank Markus Zweckstetter and his colleagues at the Max Planck Institute for Biophysical Chemistry for X-Ray Crystallography and NMR Software–National Institutes of Health discussions.

References

[1] Taylor WR, May ACW, Brown NP, Aszodi A. Protein structure: Geometry, topology and classification. Rep Prog Phys 2001; 64:517-90.

[2] Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The XPLOR-NIH NMR molecular structure determination package. J Magn Reson 2003;160:65-73.

[3] Meiler J, Baker D. Rapid protein fold determination using unassigned NMR data. Proc Natl Acad Sci USA 2003;100:15404-9.

[4] Vendruscolo M, Najmanovich R, Domany E. Protein folding in contact map space. Phys Rev Lett 1999;82:656-9.

[5] Hu J, Shen X, Shao Y, Bystroff C, Zaki MJ. Mining protein contact maps. Proceedings of the Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference); 2004 Aug 22-25.

[6] Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258:367-92.

[7] Richardson JS. Beta-sheet topology and the relatedness of proteins. Nature 1977;268:495-500.

[8] Ptitsyn OB, Finkelstein AV. Similarities of protein topologies: Evolutionary divergence, functional convergence or principles of folding. Q Rev Biophys 1980;13:339-86.

[9] Efimov AV. Structural trees for protein super-families. Proteins Struc Func Genet 1997;28:241-60.

[10] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH — a hierarchic classification of protein domain structures. Structure 1997;5:1093-108.

[11] Taylor WR. A "periodic table" for protein structures. Nature 2002;416:657-60.

[12] Demirel M, Atilgan A, Jernigan R, Erman B, Bahar I. Identification of kinetically hot residues in proteins. Protein Sci 1998;7:2522-32.

[13] Demirel M, Bahar I, Atilgan A. Predicting the ensemble of unfolding pathways for proteins: an updated incremental Lagrangean model. Biophys J 1999;76:A176.

[14] Atilgan A, Durell S, Jernigan R, Demirel M, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 2001;80:505-15.

[15] Bahar I, Atilgan A, Demirel M, Erman B. Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. Phys Rev Lett 1998;80:2733 - 6.

[16] Keskin O. Comparison of full-atomic and coarse-grained models to examine the molecular fluctuations of c-AMP dependent protein kinase. J Biomol Struc Dynamics 2002;20:333 - 46.

[17] Demirel MC, Keskin O. Protein interactions and fluctuations in a proteomic network using an elastic network model. J Biomol Struc Dynamics 2005;22:381 - 6.

[18] Flory PJ. Statistical thermodynamics of random networks. Proc R Soc London A 1976;351:351 - 80.

[19] Brunger AT, Clore GM, Gronenborn AM, Saffrich R, Nilges M. Assessing the quality of solution nuclear-magnetic-resonance structures by complete cross-validation. Science 1993;261:328 - 31.

[20] Fersht AR, Daggett V. Protein folding and unfolding at atomic resolution. Cell 2002;108:573 - 82.

[21] Mayor U, Johnson CM, Daggett V, Fersht AR. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. Proc Natl Acad Sci U S A 2001;98:13518 - 22.

[22] Ptitsyn OB. How molten is the molten globule? Nat Struc Biol 1996; 3:488 - 90.

[23] Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256:623 - 44.

[24] Baysal C, Atilgan AR. Coordination topology and stability for the native and binding conformers of chymotrypsin inhibitor 2. Proteins Struc Func Genet 2001;45:62 - 70.

[25] Humphrey W, Dalke A, Schulten K. VMD—visual molecular dynamics. J Mol Graphics 1996;14:33 - 8.