Available online at www.sciencedirect.com
SCIENCE DIRECT®
Polymer 45 (2004) 659–668
polymer
www.elsevier.com/locate/polymer

Folding core predictions from network models of proteins

A.J. Rader, Ivet Bahar*

Center for Computational Biology and Bioinformatics, School of Medicine, University of Pittsburgh, 200 Lothrop Street, W1043 Biomedical Science Tower, Pittsburgh, PA 15261, USA

Received 30 May 2003; received in revised form 25 June 2003; accepted 25 June 2003

Abstract

Two different computational methods are employed to predict protein folding nuclei from native state structures, one based on an elastic network (EN) model and the other on a constraint network model of freely rotating rods. Three sets of folding cores are predicted with these models, and their correlation against the slow exchange folding cores identified by native state hydrogen–deuterium exchange (HX) experiments is used to test each method. These three folding core predictions rely on differences in the underlying models and relative importance of global or local motions for protein unfolding/folding reactions. For non-specific residue interactions, we use the Gaussian Network Model (GNM) to identify folding cores in the limits of two classes of motions, shortly referred to as global and local. The global mode minima from GNM represent the residues with the greatest potential for coordinating collective motions and are explored as potential folding nuclei. Additionally, the fast mode peaks that have previously been labeled as the kinetically hot residues are identified as a second folding core set dependent on local interactions. Finally, a third folding core set is defined by the most stable residues in a simulated thermal denaturation procedure of the FIRST software. This method uses an all-atomic analysis of the rigidity and flexibility of protein structures, which includes specific hydrophobic, polar and charged interactions. Comparison of the three folding core sets to HX data indicate that the fast mode peak residues determined by the GNM and the rigid folding cores of FIRST provide statistically significant enhancements over random correlation. The role of specific interactions in protein folding is also investigated by contrasting the differences between these two network-based computational methods.

Keywords: Protein folding; Hydrogen exchange; GNM

1. Introduction

The theory of protein folding remains an unresolved question in modern structural biology. Various models to describe the process of how a polymer chain can reproducibly reach a unique folded conformation from an unfolded sequence of amino acids have been proposed. These models rely on the concept of a multidimensional free energy landscape, which is funnel-shaped. This concept nicely describes the way proteins reliably fold into their energetically most favorable conformations. However, because the reaction coordinates of this funnel are unknown, exact pathways rarely correspond to experimental data points. As experiments reveal more information about unfolding and folding kinetics a clearer connection between the theory and experiment is required.

So far the testing of theories has relied heavily on increasingly sophisticated computer simulations and calculations to simulate experimental results. However, even with increasing computer power, there have only been a few molecular dynamics (MD) simulations where native-like, folded structures have been obtained from an unfolded state [1,2]. These simulations are for very small sub-domains of proteins. Far more computational studies focus on the reverse process of simulated unfolding. For small, two-state folding proteins that fold reversibly, this provide information about the folding pathway(s). Many of these unfolding studies have involved time-intensive MD simulations of a single protein structure [3].

Hydrogen–deuterium exchange (HX) experiments in proteins provide local probes of internal fluctuations in the proteins. Intramolecular, hydrogen-bonded amide protons experience exchange with deuterium atoms from the bath water on a wide range of scales. Although, the fastest exchanging backbone amide protons tend to reside on the protein surface, there are significantly slower exchanging amides distributed throughout the protein that indicate a

* Corresponding author. Tel.: +1-412-648-6671; fax: +1-412-648-6676.
E-mail address: bahar@pitt.edu (I. Bahar).

0032-3861/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.10.080

'protected' residue that may be important for stability. We will focus on the exchange mechanism [4] expressed by the kinetic scheme

$$\text{closed} \underset{k_{cl}}{\overset{k_{op}}{\rightarrow}} \text{open} \xrightarrow{k_x} \text{exchanged} \tag{1}$$

where an equilibrium between the 'closed' and 'open' forms of a given amino acid is assumed. Once in the open form, the amide can exchange its proton with the solvent. When $k_{cl} \gg k_{op}$, conditions favor folding of the protein, which one can express by the observed exchange rate as

$$k_{obs} = \frac{k_{op}k_x}{(k_{cl} + k_x)} \tag{2}$$

The two limits of this kinetic exchange rate are termed EX1 and EX2. The EX1 limit describes unfolding far from physiological conditions such as high pH or high denaturant concentration. In this limit, $k_{cl} \ll k_x$, and consequently $k_{obs} = k_{op}$. The EX2 limit describes unfolding under near-physiological conditions. The observed rate constant is given by $k_{obs} = K_{eq}k_x$, using $k_{cl} \gg k_x$ using and the equilibrium constant $K_{eq} = k_{op}/k_{cl}$. The free energy cost of exchange relative that experienced by the same amino acid at solvent-exposed conditions is

$$\Delta G_{HX} = -RT \ln(k_{obs}/k_x) = -RT \ln K_{eq} \tag{3}$$

The ratio k_{obs}/k_x is also referred to as the protection factor describing how likely a given amide proton is to exchange compared to when that residue is in a random polymer or small molecule. Since the computational methods employed use the native state as the starting point, we will restrict our comparisons to EX2 data.

We extend two fast computational techniques that rely upon the native structure to infer protein unfolding nuclei and pathways, and compare these predictions with native-state HX experiments. The first technique is based on the Gaussian Network Model (GNM) of proteins [5,6]. The residues are modeled as beads subject to Gaussian fluctuations connected by elastic springs that account for chain connectivity and intramolecular interactions. GNM results were previously shown to satisfactorily reproduce the experimentally measured HX protection factors [7]. This study focused on the class of motions determined by slow modes, termed global motions because of their dominant role in controlling the collective dynamics of the protein. Another GNM study identified the so-called kinetically hot residues [8] from the peaks in the high frequency modes. These positions were pointed out to be usually conserved within protein families and linked to potential folding nuclei [8]. Both slow mode minima which indicate the hinge sites in the global motions, and fast mode peaks that are centers of localization of energy, will be tested against experimentally determined folding cores.

A different network model of proteins is the constraint network model employed in the Floppy Inclusions and Rigid Substructure Topography (FIRST) software [9,10]. In contrast to the GNM, FIRST adopts a full atomic description and considers atom specificity to identify the rigid clusters of residues that are likely to form the folding cores. We extend the FIRST folding core predictions [11] to a larger set of proteins with HX data and test the significance of the results.

Results are presented for three sets of predicted folding cores: (i) GNM slow mode minima (global motions) (G), (ii) GNM fast mode peaks (kinetically hot residues) (H), and (iii) FIRST mutually rigid folding cores (F), and all three sets are compared to experimentally defined folding cores (E). The correlations between these four sets will be presented along with the level of significance for each.

2. Models and methods

We augmented the set of 10 proteins used in the original FIRST folding core comparison [11] to include all proteins that have published HX data [12–19] for the native state in the EX2 limit. The resulting set of 29 proteins is listed in Table 1.

2.1. The Gaussian Network Model (GNM)

GNM has been used on many proteins to determine the preferred motions uniquely defined by the contact topology of residues in the native state. The protein in the GNM is modeled as an elastic network (EN). Each node represents a single residue and its mean position coincides with that of the corresponding C^{α}-atom in the PDB structure. Pairs of residues located within a cutoff distance, r_c, of 7.0 Å are assumed to be connected by elastic springs (Fig. 1(A)). This creates an EN where all connected residues interact via a harmonic potential with a uniform spring constant, γ as first proposed (at atomic scale) by Tirion [20].

The dynamics of the structure in the GNM is fully defined by the topology of contacts described by the Kirchhoff matrix $\mathbf{\Gamma}$. For a network of N interacting sites, the elements of $\mathbf{\Gamma}$ are defined as

$$\Gamma_{ij} = \begin{cases} -1 & \text{if } i \neq j \quad \text{and } \boldsymbol{R}_{ij} \leq r_c \\ 0 & \text{if } i \neq j \quad \text{and } \boldsymbol{R}_{ij} > r_c \\ -\sum_{\substack{1 \leq j \leq N \\ j \neq l}} \Gamma_{ij} & \text{if } i = j \end{cases} \tag{4}$$

where $\mathbf{R}_{ij}$ is the distance between sites i and j. $\mathbf{\Gamma}$ is simply the inter-residue contact matrix and its inverse describes the correlations between residue fluctuations in the neighborhood of the native state. The diagonal elements of the inverse, $\mathbf{\Gamma}^{-1}$, are proportional to the mean-square fluctuations $\langle \Delta \mathbf{R}_i^2 \rangle$ while the off-diagonal elements $[\mathbf{\Gamma}^{-1}]_{ij}$ refer to the cross-correlations $\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle$. The proportionality constant between $\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle$ and $[\mathbf{\Gamma}^{-1}]_{ij}$ is simply $3k_BT/\gamma$ where k_B is the Boltzmann constant and T is the temperature

Table 1
Structural properties of proteins used in this study

	Protein name	Abbrev	PDB code[a]	N[b]	Resolution (Å)
1	Apo-myoglobin	apoMb	1a6m	151	1.0
2	Barnase	Bnase	1a2p	108	1.5
3	Cytochrome *c*	Cytc	1hrc	104	1.9
4	T4 lysozyme	T4lzm	3lzm	164	1.7
5	Ribonuclease T1	RnaseT1	1bu4	104	1.9
6	α-Lactalbumin	ha-LA	1hml	123	1.7
7	Chymotrypsin inhibitor 2	CI2	2ci2	64	2.0
8	Ubiquitin	Ubq	1ubi	76	1.8
9	Bovine pancreatic trypsin inhibitor	BPTI	1bpi	58	1.1
10	Interleukin-1β	IL-1b	4ilb	151	2.0
11	Hen egg-white lysozyme	HEWL	1hel	129	1.7
12	Equine lysozyme	Eqlzm	2eql	129	2.5
13	Protein A, B-domain	pAB	1bdd	60	NMR
14	Staphyloccoccal nuclease	SNase	1stn	1361	1.7
15	Ribonuclease A	RnaseA	1rbx	124	1.7
16	Ribonuclease H	RnaseH	2rn2	155	1.5
17	Guinea pig α-lactalbumin	gpa-LA	1hfx	123	1.9
18	B1 immunoglobulin-binding domain protein G	GB1	1pga	56	2.1
19	B1 immunoglobulin-binding domain protein L	LB1	2ptl	78	NMR
20	Cardiotoxin analog III	CTX-3	2crs	60	NMR
21	Tendamistat	Tnds	2ait	74	NMR
22	Single chain antibody fragment[c]	scFv	1mcp	237	2.70
23	Human acidic fibroblast growth factor-1	hFGF-1	2afg	127	2.00
24	Cytochrome c551	pacc551	351c	82	1.60
25	Outer surface protein A	ospA	1ospO	251	1.95
26	Ovomucoid third domain	OMTKY3	1iy5	54	NMR
27	Chicken src SH3 domain	cSH3	1srm	56	NMR
28	CheY	CheY	3chy	128	1.70
29	Human carbonic anhydrase I	HCA-1	1hcb	258	1.60

[a] The Protein Data Bank identification code [59].
[b] N refers to the number of residues in the protein.
[c] In keeping with the experimental data from the previous study [12], we analyzed the folded state formed by residues 1–115 in chain L and 1–122 in chain H.

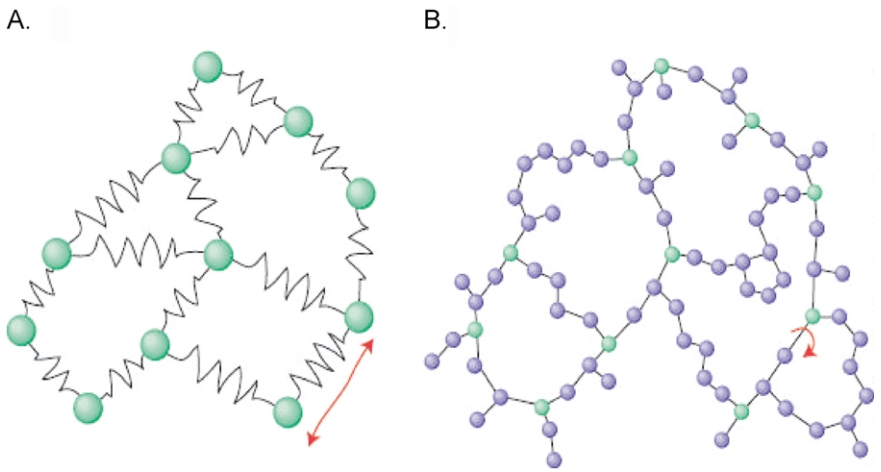

Fig. 1. Comparison of two network models for proteins. (A) The elastic network used by GNM. Every residue is represented by a single node and connected to spatial neighbors by springs. These springs then determine the N − 1 degrees of freedom in the network and the modes of vibrations about the native state. (B) The constraint network used by FIRST includes all atoms connected by fixed-length bars representing covalent bonds, hydrogen bonds, and hydrophobic interactions. (Only heavy atoms are shown in this sketch for simplicity).

[5]. A recent examination of the X-ray crystallographic B factors of over 100 proteins showed that the GNM closely reproduces the experimental data [21], using

$$\langle(\Delta\mathbf{R}_i)^2\rangle = (3k_BT/\gamma)[\mathbf{\Gamma}^{-1}]_{ii} \qquad B_i = 8\pi^2(k_BT/\gamma)[\mathbf{\Gamma}^{-1}]_{ii} \quad (5)$$

A major utility of the GNM is to calculate the shapes and frequencies of the global modes for a given quaternary structure with minimal computational cost. $N-1$ GNM modes are found by the eigenvalue decomposition of $\mathbf{\Gamma} = \mathbf{U}\Lambda\mathbf{U}^{-1}$ where $\mathbf{U}$ is the orthogonal matrix whose columns, $\boldsymbol{u}_k$, $1 \le k \le N$, are the eigenvectors of $\mathbf{\Gamma}$, and $\mathbf{\Lambda}$ is the diagonal matrix of the non-zero eigenvalues, λ_k, usually ordered in ascending order after eliminating the zero eigenvalue. The kth eigenvalue scales with the frequency of the kth GNM mode. The ith element $(\boldsymbol{u}_k)_i$ of $\boldsymbol{u}_k$ describes the motion of residue i along the kth principal coordinate. The mean-square fluctuations of individual residues can be rewritten as a weighted sum over all modes as [6]

$$\langle(\Delta\mathbf{R}_i)^2\rangle = \Sigma_k[(\Delta\mathbf{R}_i)^2]_k = \Sigma_k(3k_BT/\gamma)[\lambda_k^{-1}(\mathbf{u}_k)_i(\mathbf{u}_k)_i] \quad (6)$$

The last term in square brackets plotted against the residues index i represents the kth mode shape, i.e. the distribution of residue mobilities in the kth mode. By definition this distribution is normalized ($\Sigma_i(\mathbf{u}_k)_i(\mathbf{u}_k)_i = 1$), and λ_k^{-1} serves as a statistical weight for mode k. Slower modes thus make larger contributions to $\langle(\Delta R_i)^2\rangle$ (or B-factors). In particular, the first (slowest) modes predicted by the GNM [22–26,6], or by EN models in general [27–38], have been pointed out in numerous studies to drive domain movements relevant to biological function.

GNM global mode minima (G) are defined here as the positions whose square fluctuations in the slowest mode are less than $0.2/N$ (as opposed to $1/N$ for a uniform distribution of fluctuations among all the N residues). For the case of multi-domain proteins, the slowest mode usually describes the hinge sites at the interface between domains rather than unfolding of individual domains. Two proteins (ospA and scFv) were identified to contain more than one domain by CATH [39]. The second slowest mode was taken in these two cases, so as to possibly detect the folding core of the individual domain(s).

Residues participating in the fast modes (H) are those subject to rapid, local fluctuations. These residues are considered crucial for stability of the protein [40,41] because of their many local constraints. The peaks in the fast mode shapes, (mean-square fluctuation >0.03 for the average of the ten fastest modes) indicate the set of kinetically hot residues, H, which will be compared against the experimental set of slow exchanging residues (E).

2.2. Simulated unfolding with FIRST

The FIRST algorithm identifies rigid and flexible regions in proteins by means of a constraint network model of covalent bonds, hydrogen bonds and hydrophobic interactions [9] where nodes are connected by rigid bars instead of springs. Fig. 1(B) sketches an atomic-level constraint network for the same representative network as shown for the elastic network in Fig. 1(A) for GNM.

These nodes are subject to bond angle constraints, leaving bond rotations as the only accessible degrees of freedom. An atomic description is adopted, due to the speed of the *pebble game* algorithm, which has at worst a quadratic dependence of computing time on system size [42,9]. FIRST defines a network of constraints from an input 3D conformation of the protein and identifies each bond as flexible (rotatable) or rigid (non-rotatable). Since FIRST uses an all-atomic representation of proteins, residue specificity is inherent to the model. Within the context of the constraint network, each hydrogen bond is assigned an energy between 0 and 8 kcal/mol defined by its local geometry [10].

As in a previous comparison against a set of 10 proteins, we work under the hypothesis that the folding core is stabilized by a network of non-covalent interactions that are resistant to denaturation [11]. Thus starting from the native state, one is able to simulate thermal denaturation by removing one by one the intramolecular hydrogen bonds, in order from weakest to strongest energy. As each hydrogen bond is removed, the rigidity analysis of FIRST is recalculated providing a map of the increasing flexibility as the protein unfolds. As the protein is gradually denatured, the covalent bonds remain intact but the sizes of the rigid regions decrease and fragment reflecting an increasing flexibility with increasing temperature. Simulated thermal denaturation by breaking only hydrogen bonds in order of energy has been demonstrated to reproduce the HX unfolding pathways better than random removal of hydrogen bonds [11]. It has also been shown that all proteins undergo a rigid to flexible phase transition upon unfolding at an average coordination of 2.4 [10]. We adopt the most stable residues during the simulated unfolding as the folding core. As in the previous study, we define this set of folding core residues as the secondary structure residues that remain mutually rigid the longest in the simulated denaturation procedure. Unlike the definition of the experimental folding cores that adopt the entire secondary structures encompassing the slowest exchanging residues, we allow fractions of large secondary structures to be identified as rigid folding cores provided that at least three consecutive residues are mutually rigid with at least three consecutive residues of another secondary structural element.

2.3. Calculation of overlap

For each of the three computational methods for predicting folding nuclei, GNM fast mode peaks (H), GNM slow mode minima (G), and FIRST folding cores (F), counts of the number of residues in common between the method and the most slowly exchanging residues by HX (E) are calculated as a measure of the quality of theoretical

predictions. Likewise, the level of agreement between the computational methods is assessed by counting the number of commonly predicted residues. Calculations are repeated for each of the 29 proteins listed in Table 1.

Two quantitative measures are defined for assessing the level of agreement between methods A and B, each probing the departure from random.

The first measure, s(AB), is an *enhancement factor* defined as the ratio of the number of residues, $c(\mathrm{A},\mathrm{B})$, found in common between methods A and B to participate in the folding core, to the number expected from random selection of residues, $r(\mathrm{A},\mathrm{B})$. The second measure is the *difference*, $z(\mathrm{AB}) = c(\mathrm{A},\mathrm{B}) - r(\mathrm{A},\mathrm{B})$, between these two numbers. Let N_A and N_B denote the number of residues identified by methods A and B, respectively. The random probability $p(N_\mathrm{A}, N_\mathrm{B})$ of a match between N_A and N_B is $(N_\mathrm{A}/N)(N_\mathrm{B}/N)$, and the number of residues that will be selected by this random probability is

$$r(\mathrm{A},\mathrm{B}) = Np(N_\mathrm{A}, N_\mathrm{B}) = N_\mathrm{A}N_B/N. \tag{7}$$

The two measures of overlap between methods A and B become

$$s(\mathrm{AB}) = c(\mathrm{A},\mathrm{B})N/(N_\mathrm{A}N_\mathrm{B}) \tag{8}$$
$$z(\mathrm{AB}) = c(\mathrm{A},\mathrm{B}) - N_\mathrm{A}N_\mathrm{B}/N.$$

The values $s = 1$ and $z = 0$ correspond to random matches. The quality of agreement increases with increasing s and z.

Table 2 lists the values calculated for each of the 29 proteins analyzed in the present study. All six combinations of the four methods (three computational and one experimental) were analyzed. Mean values $\langle s\rangle$ and $\langle z\rangle$ over the 29 proteins, and their standard deviations, are given for each pair in the last two rows. The overlap parameters depend upon how many residues are selected by each method. The number of fast mode peaks is generally smaller than the number of residues selected by the experimental procedure outlined by Li and Woodward [12].

3. Results

Fig. 2 shows the folding cores predicted by each of the

Table 2
Correlation measures for folding core prediction methods

	s(EH)	s(EF)	s(FH)	s(EG)	s(GH)	s(GF)	z(EH)	z(EF)	z(FH)	z(EG)	z(GF)	z(GH)
apoMb	1.377	**3.084**	**2.157**	**2.329**	0.809	**1.812**	1.642	**16.219**	2.682	**16.550**	5.377	-0.709
bnase	**3.333**	**2.971**	**2.314**	**1.909**	**1.841**	**1.894**	**7.000**	**17.250**	**5.111**	**10.000**	**12.741**	**4.111**
cytc	**3.200**	**3.200**	**2.773**	1.231	0.800	1.333	5.500	**16.500**	**5.115**	0.750	1.250	-0.250
T4lzm	1.268	0.702	1.242	**2.161**	0.497	0.000	1.689	-5.512	0.976	**10.744**	-5.902	-1.012
Rnase T1	**4.370**	**3.294**	**2.286**	1.359	1.238	0.667	**7.712**	4.875	2.250	2.115	-1.500	1.154
ha-LA	1.491	1.435	1.118	**2.681**	**1.720**	**1.892**	1.317	4.244	0.423	**10.659**	**7.545**	1.675
CI2	1.103	0.940	1.185	0.768	1.159	1.031	0.563	-1.469	1.875	-4.844	1.188	1.375
Ubq	**1.827**	1.070	1.070	1.247	1.315	1.070	4.526	2.105	0.855	4.158	2.632	2.158
BPTI	**1.726**	**3.255**	**2.417**	1.184	**1.726**	**1.776**	2.103	7.621	4.103	0.621	2.621	2.103
IL-1b	1.348	**2.555**	**1.766**	**1.648**	**1.748**	1.177	0.775	**16.430**	3.470	4.325	2.411	2.139
hewl	**1.929**	0.896	1.034	0.860	0.794	0.215	3.372	-1.395	0.163	-0.977	-7.302	-0.519
Eqlzm	**2.092**	**1.743**	**2.419**	**1.550**	0.956	0.896	3.132	3.411	1.760	4.256	-0.349	-0.093
pAB	**1.607**	1.382	**1.935**	0.964	**1.500**	**1.742**	3.400	5.533	**5.800**	-0.333	**7.667**	2.000
Snase	**1.902**	**6.725**	0.883	0.349	**2.473**	0.000	0.949	7.662	-0.132	-1.868	-3.088	**3.574**
RnaseA	**3.000**	1.444	**2.009**	1.091	0.470	0.000	**6.000**	4.000	3.516	0.500	-6.387	-1.129
RnaseH	0.477	0.995	**1.914**	**3.307**	0.000	0.702	-1.097	-0.065	**6.206**	**11.161**	-4.677	-2.516
gpa-LA	0.447	**2.811**	0.000	**2.916**	**1.657**	**2.169**	-1.236	7.732	-1.878	**10.512**	5.390	1.585
GB1	**1.667**	1.355	**1.615**	0.602	0.718	0.000	4.804	1.571	1.143	-1.321	-0.857	-0.393
LB1	**2.182**	**2.086**	**1.765**	**1.773**	0.600	0.918	**6.500**	7.808	2.167	**6.538**	-0.359	-1.333
CTX-3	**3.195**	**1.846**	1.108	**2.517**	1.259	**2.182**	**6.183**	4.583	0.583	3.617	5.417	0.617
Tnds	**2.921**	0.974	**1.542**	1.263	1.167	0.200	5.919	-0.135	1.757	2.500	-8.000	1.000
scFv	1.484	1.467	**1.995**	1.240	1.254	0.992	1.304	9.228	2.494	**7.544**	-0.241	1.013
hFGF-1	**4.233**	**2.540**	1.411	0.977	0.543	1.465	4.583	2.425	0.291	-0.094	0.953	-0.843
pacc551	**1.621**	0.000	**1.945**	**2.228**	0.932	0.000	1.915	-6.451	2.915	**8.268**	-6.732	-0.220
ospA	**3.508**	**1.960**	**1.741**	1.364	0.000	1.312	**6.434**	**18.124**	4.255	3.470	6.661	-2.900
OMTKY3	1.142	**2.077**	1.118	1.350	1.246	0.623	5.185	1.370	0.741	2.593	-3.630	1.185
cSH3	**1.778**	**2.545**	**1.636**	1.167	1.273	1.250	5.464	7.000	3.500	0.429	1.000	0.429
cheY	1.173	1.365	**1.925**	1.102	1.407	1.080	2.672	5.164	**5.766**	2.781	1.844	2.313
HCA-I	0.701	1.020	**2.932**	0.623	**2.172**	**1.991**	0.039	-2.558	3.953	-7.256	**19.907**	**5.395**
mean	2.004	1.991	1.698	1.509	1.147	1.048	3.391	5.285	2.478	3.703	1.227	0.755
stdev	1.044	1.265	0.628	0.729	0.587	0.713	2.542	6.649	2.035	5.334	6.200	1.874

For the enhancement values, s, bold numbers emphasize greater than 50% improvement over random agreements between two methods. The bold numbers in the paired difference values, z, indicate values larger than one standard deviation from the mean value. The two measures of overlap, s and z, are discussed in the text and presented for each pair wise comparison of the results from experiments (E), and computations based on GNM global modes (G), fast mode peak residues (H), and FIRST (F).

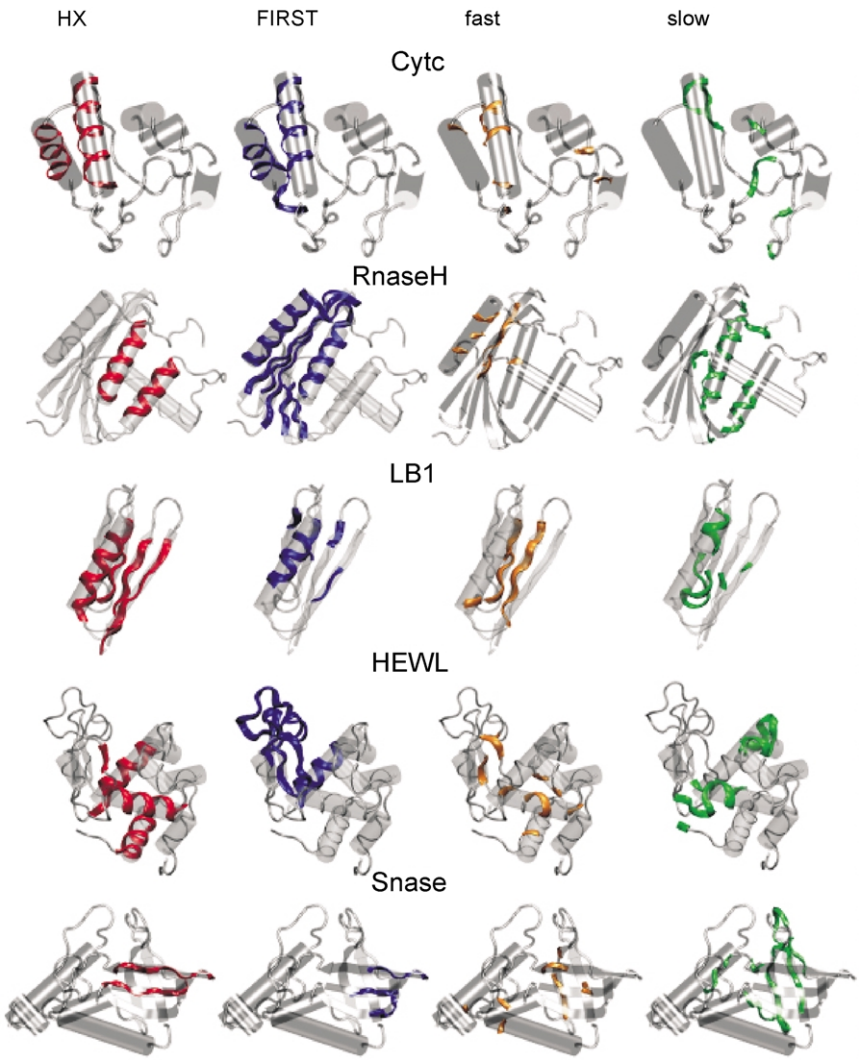

Fig. 2. Comparison of folding core predictions by four different methods. The four different folding core predictions: experimental (HX), FIRST, GNM fast mode peaks (fast), and GNM slow mode minima (slow) are mapped onto the 3D structures of five representative proteins. Helices are shown by cylinders and β-strands by arrows. The HX slow exchange folding core is shown in red, the FIRST rigid folding core is shown in blue, the GNM fast mode peaks are shown in orange and the GNM slow mode minima are shown in green. Abbreviations for the five proteins shown in this figure are taken from Table 1. The images in this figure were created using VMD [60].

four methods mapped onto the three-dimensional (3D) structures of horse cytochrome *c* (cyt *c*, PDB code 1hrc), ribonuclease H (Rnase H, 2rn2), B1 immunoglobulin-binding domain protein L (LB1, 2ptl), hen egg white lysozyme (HEWL, 1hel), and *staphyloccoccal* nuclease (Snase, 2stn). These five proteins serve as representative examples for the overlap of predicted folding cores for each of the four methods. These 3D images indicate the spatial localization of folding cores by colored ribbons and the secondary structures by gray features. Of the results in Fig. 2, the GNM slow mode minima tend to correlate the poorest with the other methods. The exception to this is for Rnase H where the slow modes actually provide the best overlap ($s(EG) = 3.3$) with experimental results. In the cases of HEWL and RnaseH, the overlap between FIRST and experiment suffer due to too many false positives.

Fig. 3 gives a reduced representation for all 29 proteins, comparing the four folding core prediction methods. In this image, the folding cores are stacked above one another and plotted versus the residue number. The colored blocks indicate residues that belong to a particular folding core set from each of the four methods: experimental, E, in red; GNM slow mode minima, G, in green; GNM fast mode peak residues, H, in orange; and FIRST, F, in blue. This qualitative comparison gives a visual representation of the consensus between methods for any given protein. The overall correspondence is quite striking for such simple native-based methods.

3.1. Comparison to HX data

We augmented the study of Li and Woodward [12] with

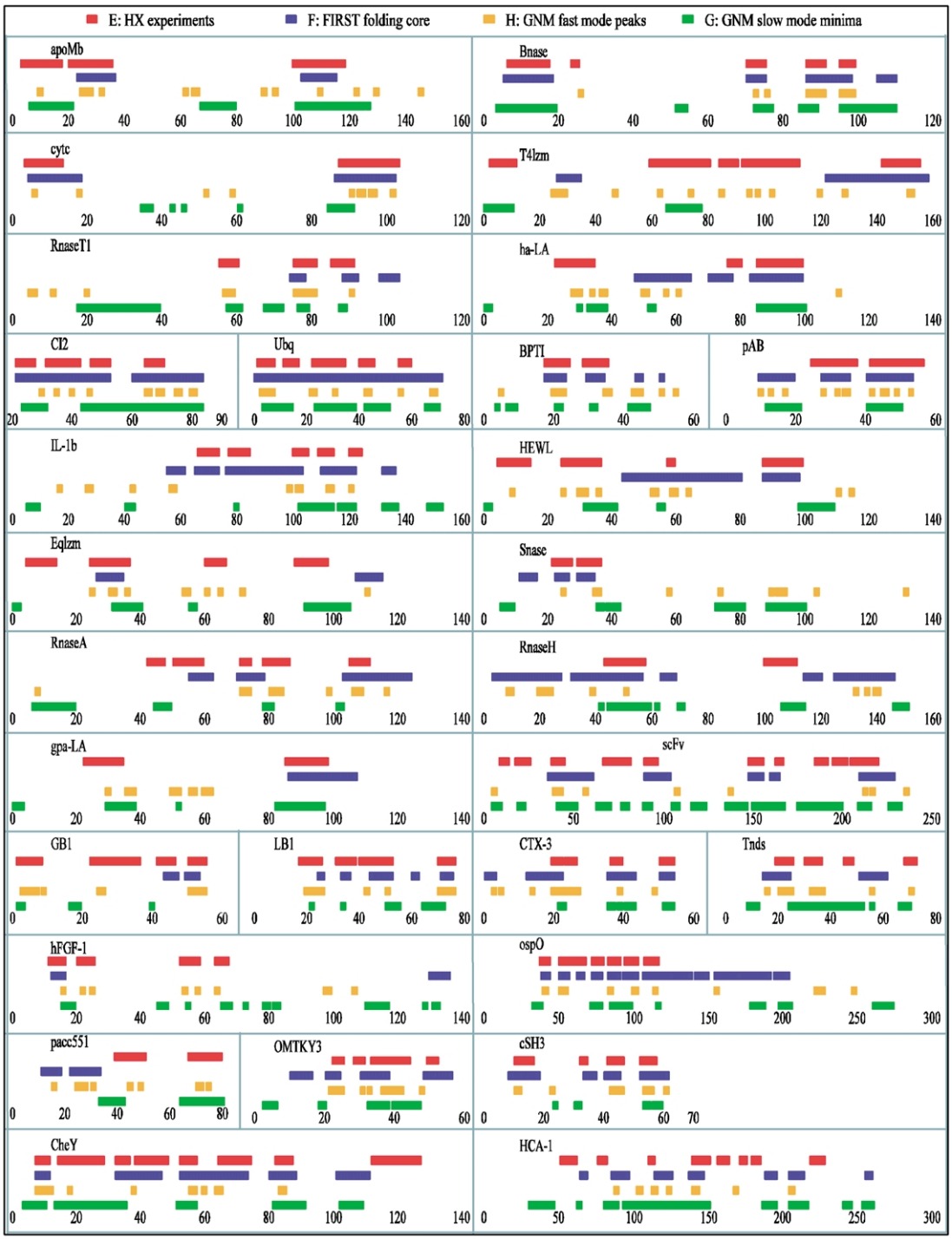

Fig. 3. Reduced representation of protein folding cores. The folding core predictions for HX slow exchange (red), FIRST rigid (blue), GNM fast mode peak (orange) and GNM slow mode minima (green) residues are plotted in lines along the sequence for each of the proteins listed in Table 1.

seven additional proteins that were unpublished at the time of their study (hFGF-1 [14], pacc551 [13], ospA [15], OMTKY3 [16], cSH3 [17], cheY [19], HCA-I [18]). Keeping with their work, we defined the experimental folding core as the secondary structural elements containing the residues with the greatest protection factors in the HX experiments. These residues became part of set E. Residues were assigned to secondary structures using the Dictionary of Secondary Structures of Proteins (DSSP) [43].

3.2. Statistical significance

Table 2 indicates that according to the enhancement factor (s) most of the correlations between pairs of methods

were at least slightly better than random ($s = 1$). The FIRST rigid cores (F) and GNM fast mode peaks (H) compared to the experimental HX slow exchange cores (E) yield on average enhancement factors twice as good at random chance, while the enhancement drops to 1.5 in the case of the comparison of the results from GNM slow modes (G) with experiments. GNM fast mode peaks are also distinguished by their small departure from the mean value, indicated by a standard deviation of 1.044 in Table 2. Although both EF and EH exhibit high enhancement factors, the correlation between these methods (indicated by FH) is much less indicating that the native state folding information extracted by each method is more complementary than redundant. This complementarity is easier to visualize by looking at Figs. 2 and 3.

Two measures for how accurate a method is at predicting observed results are sensitivity and selectivity, which depend upon the number of true positives (TP), false positives (FP) and false negatives (FN) predicted by a given method. The number of TPs is just the value of $c(A, E)$ in the present case, while the number of FPs is the number of residues selected by theory that are not part of the experimental folding core. The number of FNs is the number of residues identified by experiment as part of the folding core but not selected by the theoretical method. Sensitivity, defined as TP/(TP + FN), reflects the ability of a theoretical method to report *all* experimentally selected residues. Specificity, defined as TP/(TP + FP), complements this information with a measure of the ability to report *only* the experimentally selected residues. Fig. 4 plots the sensitivity (panel A) and specificity (panel B) of the two sets F and H that have been observed above to yield a reasonable description of HX slow exchanging core data (E) for each of the 29 proteins. The expected sensitivities from random models for these proteins are 0.13 for the pair EH and 0.33 for EF, and the expected specificity is 0.29 for both cases. In general, the majority of the data points lie above these values indicating an improvement over random assignments. For any individual protein, the difference between the two points indicates how the two methods differ in their performance. FIRST is distinguished by a higher sensitivity, while GNM fast modes appear to be slightly more specific in general.

To test if these departures were significant from random selections of residues, the paired difference values, z, were calculated. The appropriate statistical analysis for such nonparametric data is the Wilcoxon signed rank test [44]. According to this statistical analysis, the null hypothesis that selecting a given number of residues in common by two methods is no better than random selection may be rejected because the probability of such an event is less than 0.0005 for the cases of EH ($|z^-| = 4.465$), EF ($|z^-| = 3.535$), and HF ($|z^-| = 4.422$), where $|z^-|$ is the departure from the mean for a standard distribution centered at 0. Thus the correlations indicated by large values for s(EH), s(EF) and s(HF) indicate that these departures are better than random selection. This indicates a strong likelihood that the fast mode peaks and FIRST folding core methods accurately predict the folding nuclei. Applying the same analysis to the other pairs of data (EG, HG, and FG) resulted in much lower

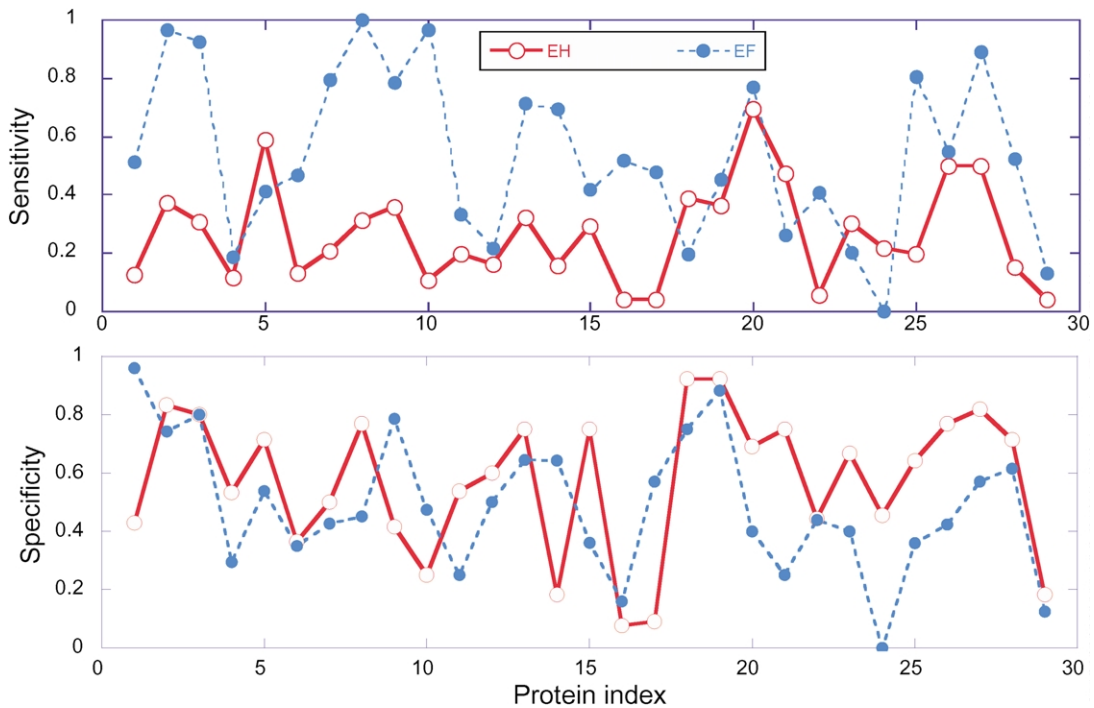

Fig. 4. Sensitivity (A) and specificity (B) analysis of FIRST and GNM fast modes predictions with respect to the HX slow exchanging cores for all 29 proteins listed in Table 1. In both panels, the sensitivities/specificities of the sets F and H are compared. In both panels, open circles refer to the GNM fast mode peaks, and filled circles refer to FIRST results. The FIRST folding core shows the highest sensitivity while the GNM fast modes appear to be relatively more specific.

values of $|z^-|$ and hence much higher probabilities that such correlations were the result of random chance. Figs. 2 and 3 present graphical representations of the correlations to experimental slow exchange cores indicated by these statistical methods.

4. Discussion

In recent years, there has been a revival in the number of studies that exploit Go-like models for understanding protein folding kinetics (see the introduction of Kaya and Chan's paper for reference [45]). The major assumption in these studies is that the native state contact topology plays a dominant role in defining the protein folding kinetics. The validity and utility of such coarse-grained approaches have been tested on numerous protein systems and applications. While many features relevant to folding kinetics, particularly in the case of fast-folding proteins, can be predicted by such topology-based models, more detailed systematic studies are needed.

The present work was undertaken to explore the predictive ability of two simple models, one based on an elastic network theory of random networks [46], and the other on a constraint network composed of rigid but rotatable connectors similar to the classical freely rotating chain model of polymer statistics [47,48].

While both models are conceptually simple, they bear exact solutions and can be readily applied to a series of proteins. The first relies on fundamental concepts of statistical thermodynamics, and takes rigorous account of the overall coupling of all residues. The second has the additional advantage of incorporating specificity and atomic details through a network of hydrogen bonds and hydrophobic contacts. Fig. 1 shows the same small network for both the elastic network (A) and the constraint network (B). Our analysis demonstrates that either method can be utilized for predicting to a reasonable accuracy level the slow exchanging residues observed in HX. Whether these residues form the folding nuclei is another issue, but to the extent that they do, the present topology-based coarsegrained models give insights about the identity or location of folding nuclei.

Correlations between experimental ϕ-value analysis which has been used to identify potential folding nuclei from the most stable substructures formed in the transition state, and these native state folding core predictions were not made. Since, the ϕ-value analysis relies upon observing how mutations disrupt crucial non-local interactions and the resulting folding rates, these ϕ-values relate more directly to contacts made at the transition state than those forming the folding cores; and may not be used directly to interpret the degree of the folded structures.

The relative contact order has been shown to correlate with the folding rates for small two-state protein folders [49]. This theory implies that native-state topology has a dominant role in the rates of protein folding, at least for two-state folders. A recent study of multi-state protein folders has indicated that there is, however, no correlation between the folding rate and relative contact order [50]. Such contradictory results indicate that the complexity of the protein folding problem and the need to explore folding using methods that explore different reaction coordinates.

Several groups have employed GNM methods to explore protein folding. One study created a minimization methodology for a simplified lattice model of proteins involving only hydrophobic and polar residues (HP) based upon a native state Gaussian fluctuations [51,52]. Although elegant in the time evolution of states, application to a set of five real proteins has reproduced the native folded state of these proteins with root mean square deviations greater than 3.5 Å [52]. A second approach introduces temperature into the temperature-free GNM representation as a measure of the probability of interactions between residue pairs. With this approach, the native state fluctuations are used as a starting point to investigate the unfolding as it is allowed to depart from the native-state biased by the global vibrational modes [53–55]. This approach seems promising for exploring the HX data in the EX1 limit (far from native state conditions), which is beyond the scope of the present study.

Computational comparisons to experimental HX data to elucidate protein folding have been conducted previously. A study that selected folding nuclei only on proximity to charge centers suggested that electrostatics play an important role in determining protein folding [56]. However, charged residues exhibit a relatively high probability to cluster around the active sites of proteins [57] and it is likely that the nuclei selected by such an electrostatics method are active-site residues rather than folding nuclei. Other computational methods of actually reproducing experimental protection factors by generating a large ensemble of partially folded states [58] require much greater computational effort, and previous comparison with experimental data and GNM predictions does not lend support to the adoption of such expensive computations.

The results presented here are for four specific methods for selecting folding nuclei. Other selection schemes could be devised as well as better ways to optimize the selection processes. However, the goal of this study was to investigate the probability of selecting the residues critical for folding from simple network models of the native state. Even with these simple and fast computational methods, the claim that native state fluctuations and flexibility encode for protein folding is strongly supported. We assume the native state contains information concerning how the protein folded into that native state. The study here used two approaches based upon the native structure information to investigate the folding nuclei. The complementarity of these methods is indicated by the fact that although both FIRST and GNM fast modes can accurately predict the folding nuclei, the enhancement factor between these methods is only moderate, indicated by a value of 1.698.

References

[1] Duan Y, Wang L, Kollman PA. Proc Natl Acad Sci 1998;95: 9897–902.
[2] Chowdhury S, Lee MC, Xiong G, Duan Y. J Mol Biol 2003;327: 711–7.
[3] Daggett V, Li A, Itzhaki LS, Otzen DE, Fersht AR. J Mol Biol 1996; 257:430–40.
[4] Hvidt A, Nielsen SO. Adv Protein Chem 1966;21:288–386.
[5] Bahar I, Atilgan AR, Erman B. Folding Des 1997;2:173–81.
[6] Haliloglu T, Bahar I, Erman B. Phys Rev Lett 1997;79:3090–3.
[7] Bahar I, Wallqvist A, Covell DG, Jernigan RL. Biochemisty 1998;37: 1067–75.
[8] Demirel MC, Atilgan AR, Jernigan RL, Erman B, Bahar I. Protein Sci 1998;7:2522–32.
[9] Jacobs DJ, Rader AJ, Thorpe MF, Kuhn LA. Proteins 2001;44: 150–65.
[10] Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF. Proc Natl Acad Sci 2002;99:3540–5.
[11] Hespenheide BM, Rader AJ, Thorpe MF, Kuhn LA. J Mol Graph Modell 2002;21:195–207.
[12] Li R, Woodward C. Protein Sci 1999;8:1571–91.
[13] Russell BS, Zhong L, Bigotti MG, Cutruzzola F, Bren KL. J Biol Inorg Chem 2003;8:156–66.
[14] Chi Y-H, Kumar TKS, Kathir KM, Lin D-H, Zhu G, Chiu I-M, Yu C. Biochemistry 2002;41:15350–9.
[15] Yan S, Kennedy SD, Koide S. J Mol Biol 2002;323:363–75.
[16] Arrington CB, Teesch LM, Robertson AD. J Mol Biol 1999;285: 1265–75.
[17] Grantcharova VP, Baker D. Biochemistry 1997;36:15685–92.
[18] Kjellsson A, Sethson I, Jonsson BH. Biochemistry 2003;42:363–74.
[19] Lacroix E, Bruix M, Lopez-Hernandez E, Serrano L, Rico M. J Mol Biol 1997;271:472–87.
[20] Tirion MM. Phys Rev Lett 1996;77:1905–8.
[21] Kundu S, Melton JS, Sorensen Jr. DC. Biophys J 2002;83:723–32.
[22] Bahar I. Rev Chem Eng 1999;15:319–47.
[23] Bahar I, Atilgan AR, Demirel MC, Erman B. Phys Rev Lett 1998;80: 2733–6.
[24] Bahar I, Erman B, Jernigan RL, Atilgan AR, Covell DG. J Mol Biol 1999;285:1023–37.
[25] Bahar I, Jernigan RL. J Mol Biol 1998;281:871–84.
[26] Bahar I, Jernigan RL. Biochemistry 1999;38:3478–90.
[27] Hinsen K. Proteins 1998;33:417–29.
[28] Hinsen K, Thomas A, Field MJ. Proteins 1999;34:369–82.
[29] Delarue M, Sanejouand YH. J Mol Biol 2002;320:1011–24.
[30] Elezgaray J, Sanejouand YH. Biopolymers 1998;46:493–501.
[31] Keskin O, Bahar I, Flatow D, Covell DG, Jernigan RL. Biochemistry 2002;41:491–501.
[32] Keskin O, Durell SR, Bahar I, Jernigan RL, Covell DG. Biophys J 2002;663–80.
[33] Keskin O, Ji X, Blaszcyk J, Covell DG. Proteins 2002;49:191–205.
[34] Tama F, Brooks III CL. J Mol Biol 2002;318:733–47.
[35] Tama F, Wriggers W, Brooks III CL. J Mol Biol 2002;321:297–305.
[36] Temiz NA, Bahar I. Proteins 2002;49:61–70.
[37] Ming D, Kong Y, Lambert MA, Huang Z, Ma J. Proc Natl Acad Sci 2002;99:8620–5.
[38] Ming D, Kong Y, Wakil SJ, Brink J, Ma J. Proc Natl Acad Sci 2002; 99:7895–9.
[39] Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. Nat Struct Biol 2000;structural genomics supplement:991–4.
[40] Demirel MC, Atilgan AR, Jernigan RL, Erman B, Bahar I. Protein Sci 1998;7:2522–32.
[41] Bahar I, Atilgan AR, Demirel MC, Erman B. Phys Rev Lett 1998;80: 2733–6.
[42] Jacobs DJ, Thorpe MF. Phys Rev Lett 1995;75:4051–4.
[43] Kabsch W, Sander C. Biopolymers 1983;22:2577–637.
[44] Schmidt MJ. Understanding and using statistics, 2nd ed. Lexington, Massachusetts: D.C. Heath and Co; 1979.
[45] Kaya H, Chan HS. J Mol Biol 2003;326:911–31.
[46] Flory PJ. Proc R Soc Lond A 1976;351:351–80.
[47] Flory PJ. Statistical mechanics of chain molecules, 2nd ed. New York: Hanser Publishers; 1989.
[48] Mattice WL, Suter UW. Conformational theory of large molecules. The rotational isomeric state model in macromolecular systems. New York: Wiley; 1994.
[49] Plaxco KW, Simons KT, Baker D. J Mol Biol 1998;277:985–94.
[50] Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV. Proteins 2003;51:162–6.
[51] Erman B, Dill K. J Chem Phys 2000;112:1050–6.
[52] Erkip A, Erman B, Seok C, Dill K. Polymer 2002;43:495–501.
[53] Micheletti C, Banavar JR, Maritan A. Phys Rev Lett 2001;87:3372–5.
[54] Banavar JR, Maritan A. Proteins 2001;42:433–5.
[55] Micheletti C, Lattanzi G, Maritan A. J Mol Biol 2002;321:909–21.
[56] Torshin IY, Harrison RW. Proteins 2001;43:353–64.
[57] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. Nucl Acids Res 2002;30:276–80.
[58] Hilser VJ, Freire E. J Mol Biol 1996;262:756–72.
[59] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. Nucl Acids Res 2000;28:235–42.
[60] Humphrey W, Dalke A, Schulten K. J Mol Graph 1996;14:33–8.