Elastic network normal modes provide a basis for protein structure refinement

Pawel Gniewek,[1,2] Andrzej Kolinski,[1] Robert L. Jernigan,[3] and Andrzej Kloczkowski[2,3,4,a)]
[1]*Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland*
[2]*Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, 700 Children's Drive, Columbus, Ohio 43205, USA*
[3]*Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011-3020, USA*
[4]*Department of Pediatrics, The Ohio State University, Columbus, Ohio 43205, USA*

(Received 22 February 2012; accepted 18 April 2012; published online 18 May 2012)

It is well recognized that thermal motions of atoms in the protein native state, the fluctuations about the minimum of the global free energy, are well reproduced by the simple elastic network models (ENMs) such as the anisotropic network model (ANM). Elastic network models represent protein dynamics as vibrations of a network of nodes (usually represented by positions of the heavy atoms or by the C^{α} atoms only for coarse-grained representations) in which the spatially close nodes are connected by harmonic springs. These models provide a reliable representation of the fluctuational dynamics of proteins and RNA, and explain various conformational changes in protein structures including those important for ligand binding. In the present paper, we study the problem of protein structure refinement by analyzing thermal motions of proteins in non-native states. We represent the conformational space close to the native state by a set of decoys generated by the I-TASSER protein structure prediction server utilizing template-free modeling. The protein substates are selected by hierarchical structure clustering. The main finding is that thermal motions for some substates, overlap significantly with the deformations necessary to reach the native state. Additionally, more mobile residues yield higher overlaps with the required deformations than do the less mobile ones. These findings suggest that structural refinement of poorly resolved protein models can be significantly enhanced by reduction of the conformational space to the motions imposed by the dominant normal modes. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4710986]

I. INTRODUCTION

Accurate prediction of protein three-dimensional structures from the amino acid sequence is an ultimate goal in computational molecular biology.[1] In the last three decades we have seen significant progress in this field; however, it still remains one of the most important challenges, especially for proteins having a low degree of sequence similarity to known proteins structures. The importance of the protein structure prediction problem reflects the key role of proteins in living organisms, and potential significance of structure predictions in the pharmaceutical and *biotech* industries. A more accurate knowledge of protein structure is a critical basis for a rational, computer-aided drug design.[2,3] This deeper structural knowledge can significantly enhance the prediction and understanding of the mechanisms of protein function.[4] However, with the high cost of experimental protein structure determination (estimated in 2008 to be on average $60 000–70 000 per single Protein Data Bank (PDB) entry),[5] gains from computational approaches are generally perceived to be important, as evidenced by the National Institutes of Health Protein Structure Initiative (PSI) homology modeling program. More important, a better understanding of protein function requires prior knowledge of its structure. Usually protein function prediction follows a traditional scheme assuming that the protein sequence determines its structure, and a static structure determines its function.[6] Despite a conventional static view of protein structure based on crystallographic data, certain parts of protein structure, especially those outside the protein core are relatively mobile.[7] All atoms in the structure fluctuate around their mean positions with some fluctuations being as large as 10 Å.[8] However, these motions are not independent, and often depend on the entire structure, an observed high level of cooperativity is well captured by the elastic network models. These thermal fluctuations are manifested in the crystallographic Debye-Waller temperature factors (B-factors), although often confounded by intermolecular interactions. It is well established that protein function often cannot be fully explained by its static structure only.[9–11] The dynamics is observed directly in multiple structures of the same, or similar proteins, as revealed by principal component analysis (PCA),[12] which shows close agreement with the elastic network models (ENMs).[8,13–21] The ENMs have proven their ability to sample efficiently conformational space. Especially noteworthy has been their ability to agree with the conformational transition direction between two forms of the same protein. These studies by us and others provide important evidence that structures, even when coarse-grained,

a)Author to whom correspondence should be addressed. Electronic mail: Andrzej.Kloczkowski@nationwidechildrens.org.

have an intrinsic tendency to fluctuate in limited numbers of directions.[22–29] The present work to refine predicted protein structures is built upon this foundation.

Other experiments such as NMR,[30] site-directed spin labeling (SDSL)[31,32] and hydrogen exchange can also provide dynamics information. All these studies provide deeper insights into protein physics, which lead to improved methods for protein structure prediction, protein engineering, and rational drug design. However, most researchers focus on the dynamics and the structure of biomacromolecules near the global free-energy minimum (native state). Despite the fact that during the process of folding or misfolding (both *in vivo* and *in vitro*) protein traverses through many intermediate states,[33–35] much less attention is being paid to the structures in non-native (intermediate) states. The large number of local (non-global) energy minima predicted by the theory of the protein energy landscape,[36,37] makes it practically impossible to analyze all possible non-native states. Usually for a set of predicted structures, energy minimization is performed to obtain the structure corresponding to the closest energy minimum to the native one.[38] Such an approach usually requires a full-atomic representation of the system. Additionally, the reliability of such approaches can be questioned because of the difficulty of seeking the global energy minimum.[39,40] There is not a reliable energy function which, for a given protein, can distinguish the native state from all non-native conformations. Most of the time a structure will reside within the global energy minimum, but thermal interactions with solvent can result in energy exchanges, conformational transitions, and even unfolding.[35] There is also a non-zero probability that a protein in local, non-native state energy minimum can traverse energy barriers, and reach the global energy basin. Such a model of protein dynamics in the local state was proposed by Kitao *et al.*, and called the jumping-among-minima (JAM) model.[41] In this model, protein motions in a state close to the native one are divided into two groups: (i) Harmonic modes of higher frequency, which sample a basin of each substate, and are responsible for intrastate vibrations. (ii) Anharmonic motions which are responsible for the traversing of substates, and correspond to the slowest modes. They constitute about 5% of all motions, and are called hierarchical modes. Despite the modes responsible for large scale conformational changes being anharmonic it has been shown that for short time scales they can be represented by harmonic vibrations.[42] The situation changes for longer time scales, with the directions of vibrations changing only slightly.[42,43] Because these anharmonic motions account for the larger part of protein dynamics a "reaction" pathway between two states should be minimized.[44] Obtaining such pathway is difficult, but the elastic network models provide some assistance in this direction.[45,46]

To investigate protein function and dynamics accurate protein structures are required. Often even the most sophisticated and successful structure prediction methods (e.g., I-TASSER,[47] CABS,[48] ROSETTA[49]), cannot provide the target structures with sufficiently high accuracy. Therefore, the future progress in computational biology critically depends on the successful refinement of models generated using standard template(s)-based (or template-free) modeling techniques. The importance of structure refinement has been recently emphasized and since the 8th edition of the Critical Assessment of protein Structure Prediction (CASP) event (CASP8) a new category of refinement of protein models was proposed. The latest analyses of CASP results[50,51] suggest that the solution to this problem is still missing.

Protein modeling efforts cannot yet provide predicted models that agree perfectly with experimentally determined structures, but often these structures are close enough to warrant efforts to refine them. Predicted structures are usually obtained by clustering structures from folding simulations. It can be assumed, that if a structure from the clustering, is not in the global minimum it occupies one of the local energy minima (substates) which are closely proximate to the global one. The JAM model mentioned earlier suggests, that if one considers a set of structures in such a substate, the directions of thermal motions, even in the harmonic approximation, should overlap with the direction of deformation of structure towards the native state.[52] This expectation is additionally supported by recently performed MD simulations which demonstrate the validity of the JAM model on larger spatio-temporal scales, which serve to justify approximating anharmonic oscillations by harmonic terms.[67,68] Such an assumption is valid provided that the structure modeled has at least a moderate accuracy. In addition, it is important that the ensemble of structures comes from template-free modeling, since when a template is used, the generated ensemble of structures is artificially shifted toward the local minimum imposed by template.

The idea of using normal modes for refinement of protein structural models obtained from x-ray crystallography has been presented previously.[66] Here, we focus and test a similar concept for the purpose of refining and improving predicted protein structures. To choose such conformations, putatively in non-global energy minima, protein structures are taken from conformation decoy sets generated by I-TASSER.[47] Hierarchical structure clustering is then performed.[53] The equilibrium dynamics of these structures located at either local or global minima is described by anisotropic network model (ANM).[54] Prior to applying the ANM, all structures have been energy minimized. The overlap between the thermal motions and the direction of the deformation towards the native state is then analyzed. From the viewpoint of the JAM model these should be related, especially if one considers an ensemble of putatively near-native structures instead of only one structure.[41] We measure the overlap as the cumulative overlap (COV) between the deformation matrix (which describes how a given conformation differs from the native structure) of the system, and the set of lowest frequency normal-modes. Interestingly, a quite high correlation is observed. Correlations between the contribution of each residue to the COV and its mobility are examined. At the end, the properties of the COVs and some simple descriptors of protein structures are analyzed.

II. METHODS

A. Deformation vector

We represent a structure by its C^{α} trace where the chemical identities of the amino acids are neglected. Thus a

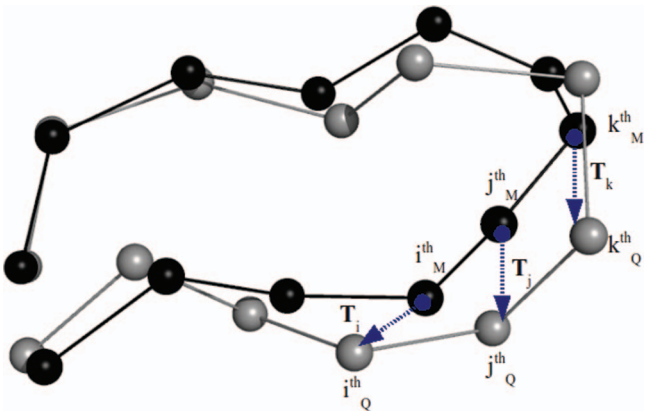

FIG. 1. Scheme of construction of the deformation matrix between two structures. Structures are superimposed, and then the *i*-th element of deformation matrix **T** is taken as a difference between the coordinates of two corresponding positions. On the picture **Q** indicates query, and the **M** modeled structure so that $\mathbf{T} = \mathbf{Q} - \mathbf{M}$.

structure can be expressed as an N × 3 coordinate matrix, where N is the number of amino-acids in the protein, and the position of the *i*-th residue is described by a set of three coordinates (x_i, y_i, z_i). Following the structural superposition of two structures, the deformation vector is calculated. If **Q** is the coordinate matrix of the native protein structure, and **M** is the coordinate matrix of the structure being compared then the deformation vector is given by

$$\mathbf{T} = \mathbf{Q} - \mathbf{M} \tag{1}$$

with **T** being the N × 3 matrix that describes the positional difference between the structures **Q** and **M**. Then **T** is resized from N × 3 to 3N × 1. An example of the construction of such a matrix is depicted in Figure 1.

B. Decoy clustering – hierarchical clustering of protein models (HCPM)

In order to condense the information from trajectories and to find putative structures in global or local energy minima, HCPM[53] is performed. In the initial step, each structure forms a separate cluster (size of one). During each iteration of the algorithm, the two closest clusters are identified and combined into a single cluster. Then program stores the identities of the combined clusters and the distances between them. The distance between any two clusters is calculated as the smallest root-mean-square deviation (RMSD) between all pairs of members of the clusters. The RMSD cut-off separating two clusters is adjusted automatically during clustering. The minimum size of retained clusters is 20 members. The clusters are sorted by the number of the members in the cluster. The most populated cluster is called the top one.

C. Datasets from I-TASSER simulation

For this study 56 trajectories have been considered. They come from *template-free* protein folding simulations carried out by using one of the most accurate algorithms for protein structure prediction: I-TASSER.[37] Each structure dataset is prepared from a long template-free simulation. Each original decoy set contains 12 500–32 000 of protein decoys. We use trajectories that are freely available from the Zhang Lab web-site (http://zhanglab.ccmb.med.umich.edu).[47,55] Details of the I-TASSER algorithm can be found elsewhere.[47] Here, we are dealing with structures considered to be at local energy minima, and near the global free energy minimum (native structure). To select possible structures trajectory clustering is performed. An appropriate structure in the local (global) energy minimum is found by taking the central member of the cluster. Such a medoid can sometimes have some structural clashes. To remove these we perform molecular mechanics minimizations in vacuum with the OPLS-AA[56] force field using GROMACS version 4.0.7.[57] Minimization stop criteria was the presence of maximal force <1 kJ mol^{-1} nm^{-1}. Electrostatic and van der Waals interactions were applied with a cut-off of 1 nm.

D. Protein quality predictor (ProQ)

For the quality assessment of each medoid the neural network method ProQ[58] was applied. This method combines the descriptors for a given structure and returns two values: LG-score and Max-Sub, which taken together describe the quality of the structure. LG-score (Max-Sub) scores higher than 1.5(0.1), 3(0.5), and 5(0.8) correspond to correct, good, and very good models, respectively. The descriptors used are non-hydrogen atom contacts with a cut-off of 5.0 Å, and residue-residue contacts with a 7.0 Å cut-off. The residues are expressed in a six letter alphabet. In addition, solvent accessibility surface area for each amino-acid and the compatibility of the secondary structure with the secondary structure predicted by PSIPRED[59] are taken into account.

E. Anisotropic network model

The anisotropic network model[54] is a coarse-grained model for the protein dynamics, which is used to investigate and describe the fluctuational dynamics in terms of the normal modes of an elastic network. It is assumed that each amino acid is represented by one point (in our case by the C^{α} atom). Then each pair of residues separated less than a cutoff distance is assumed to be connected by a harmonic spring.

If the native structure corresponds to the energy minimum, then we can expand the potential energy in the Taylor series in terms of small positional deviations of residues $\Delta\mathbf{R}$ from their mean equilibrium values $\mathbf{R}_0$

$$V(\mathbf{R}_0 + \Delta\mathbf{R}) = V(\mathbf{R}_0) + \frac{\partial}{\partial \mathbf{R}} V(\mathbf{R})_{\mathbf{R}_0} \cdot \Delta\mathbf{R} + \frac{1}{2}\frac{\partial^2}{\partial \mathbf{R}^2} V(\mathbf{R})_{\mathbf{R}_0} \cdot \Delta\mathbf{R}^2 + \ldots . \tag{2}$$

For small deviations one can neglect terms higher than the second. Taking $V(\mathbf{R}_0) = 0$, and noticing that the first derivative is equal to zero, we can rewrite that equation as

$$V(\mathbf{R}_0 + \Delta\mathbf{R}) = \frac{1}{2}\Delta\mathbf{R}^T \mathbf{H} \Delta\mathbf{R}, \tag{3}$$

where **H** is the matrix of second derivatives. If the distance between two residues i and j is less than or equal to a cutoff value (taken in this study as 18 Å),[65] then we calculate $H_{3i+k,\,3j+l}$ ($k,\,l = 1$–3) by applying a harmonic potential in the computation of the second derivative; otherwise $H_{3i+k,\,3j+l}$ is set to 0.

The Hamiltonian of the system can be written as

$$\mathbf{M}\Delta\ddot{\mathbf{R}} + \mathbf{H}\Delta\mathbf{R} = 0. \tag{4}$$

From the assumption of the harmonic nature of oscillations, we can express the Hessian as

$$\mathbf{H} = \mathbf{U}\Lambda\mathbf{U}^T, \tag{5}$$

where **U** matrix is a matrix of eigenvectors, and **Λ** is a diagonal matrix of eigenvalues. There are 3N-6 non-zero eigenvalues. The six zero eigenvalues arise from the rigid body translations and rotations corresponding to six degrees of freedom.

From such decomposition one can easily calculate the thermal fluctuations of each element of the elastic network, which can be expressed as a B-factor with the following formula

$$B_i = \frac{k_B T}{\gamma}\sum_{k=1}^{3} H^{-1}_{3i+k,3i+k}. \tag{6}$$

Here γ is the universal spring constant used in the calculation of the Hessian, k_B is the Boltzmann constant, T is temperature, and $H^{-1}_{3i+k,3i+k}$ is the $(3i+k)$-th diagonal element of the inverse of the Hessian matrix $\mathbf{H}^{-1}$.

F. Cumulative overlap

The overlap (alignment) O_i between the deformation matrix **T**, and the i-th normal mode $\mathbf{u}_i$ is defined as[44]

$$O_i = \frac{|\mathbf{T}\cdot\mathbf{u}_i|}{\|\mathbf{T}\|\cdot\|\mathbf{u}_i\|}, \tag{7}$$

where $\|\mathbf{X}\|$ denotes the norm of the vector **X**. Perfect match between the two directions occurs when $O_{\mathbf{T},i}$ is equal to 1.0. Next, we define the k-th cumulative overlap COV(k) between first k normal modes and the deformation matrix as[8]

$$COV(k) = \left(\sum_{i=1}^{k} O_i^2\right)^{1/2}. \tag{8}$$

Positional overlap between deformation matrix **T**, and i-th normal mode is defined as

$$O_{i,j} = \frac{|\mathbf{T}\cdot\mathbf{u}_i|}{\|\mathbf{T}\|\cdot\|\mathbf{u}_i\|} - \frac{|\mathbf{T}\cdot\mathbf{u}_i^{0,j}|}{\|\mathbf{T}\|\cdot\|\mathbf{u}_i\|}, \tag{9}$$

where **T**, **u**, and i have the same meaning as in Eqs. (7) and (8), and the superscript 0, j denotes that the j-th element of the $\mathbf{u}_i$ mode was set to zero. Cumulative positional k-th overlap COV(j,k) is defined by analogy to COV(k) as

$$COV(j,k) = \left(\sum_{i=1}^{k} O_{i,j}^2\right)^{1/2}. \tag{10}$$

G. Spearman correlation coefficient

Spearman's rank correlation coefficient is a non-parametric measure of the statistical dependence between two ranked variables. In the case of no tied ranks the Spearman correlation coefficient can be computed from a simple formula:

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{11}$$

with $d_i = x_i - y_i$ being the difference between the ranks of the two variables. Spearman's rank correlation coefficient is especially useful when we predict a monotonic dependence between two variables. In the case of the existence of equal rankings ρ_S is computed from the same formula as Pearson correlation coefficient (but after replacing random variables by their rankings),

$$\rho_P = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \tag{12}$$

H. TM-score

Template Modeling score (TM-score)[64] is a parameter to measure proteins structure similarity and it is defined as

$$TM-score = \frac{1}{L_Q}\sum_{1}^{L_Q}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}, \tag{13}$$

where L_Q is the length of the query protein, d_i is the Cartesian distance of two corresponding amino acids after structures superposition, and d_0 is estimated from following formula:

$$d_0 = 1.24\sqrt[3]{L_Q - 15} - 1.8. \tag{14}$$

III. RESULTS AND DISCUSSION

A. Trajectories clustering summary

Structure clustering of 56 decoys sets obtained from *template-free* protein structure predictions by I-TASSER[37] was performed. Characteristic representatives of highly populated clusters called medoids are obtained. Only cases with sufficiently large population of clusters are analyzed. Here 53 cases were considered. Table I summarizes the results of clustering. Proteins are small to medium size, with a median number of residues equal to 76. Proteins are modeled moderately well with an average RMSD of about 5.36 Å if the best structure is from the top cluster. The accuracy of prediction is increased (to RMSD = 4.55 Å) by relaxing the considerations and permitting the choice of the best structure from among the first top five clusters. Because only models close to the native structures are considered, they should have at least the same fold. As a cut-off a TM-score greater than or equal to 0.5 that is equivalent, more or less to 4.5 Å is used.[60] Following application of this restriction, 28 cases remain (from initial 53

TABLE I. Summary of clustered output. First column is the PDB name of the protein. Italic cases were not considered in later analyses because of poor resolution. The second column is the number of amino acids in the protein. The third column shows the RMSD of the medoid of the first cluster, while the fifth column shows the RMSD for the best medoid among the first top 5 clusters, and the number of that cluster. The fourth and sixth columns show the changes of RMSD after energy minimization. The superscript ^{m} denotes a median from the column.

Name	Length	Top 1 – RMSD[Å]	ΔRMSD[Å] ∼ energy minimization	Top 5 – RMSD[Å]/C. No.	ΔRMSD[Å] ∼ energy minimization
1abv	103	13.75	0.56	9.03/4	0.30
1af7	72	4.83	0.19	4.26/3	0.61
1ah9_	63	3.42	0.43	3.42/1	0.43
1b4bA	71	5.74	0.59	5.37/4	0.14
1b72A	49	2.91	0.43	2.91/1	0.43
1bm8	99	10.26	0.18	7.06/3	0.41
1bq9A	53	8.56	0.51	6.67/2	0.53
1cewI	108	4.16	0.40	3.82/2	0.46
1cqkA	101	2.08	0.52	1.91/4	0.48
1csp_	67	2.64	0.36	2.54/4	0.37
1dcjA	73	9.63	0.36	9.63/1	0.36
1di2A	69	3.59	0.42	3.06/5	0.42
1dtjA	74	2.37	0.45	1.80/4	0.46
1egxA	115	2.79	0.54	2.03/3	0.49
1fadA	92	3.64	0.49	3.35/3	0.49
1fo5A	85	3.89	0.13	3.71/4	0.54
1g1cA	98	2.68	0.37	2.49/4	0.44
1gjxA	77	8.7	0.50	8.62/5	0.44
1gnuA	117	9.81	0.54	9.44/4	0.55
1gpt	47	6.56	0.54	6.52/3	0.53
1gyvA	117	3.62	0.49	3.60/2	0.54
1hbkA	89	3.95	0.56	3.70/2	0.50
1itpA	68	10.6	0.33	10.49/3	0.01
1jnuA	104	3.27	0.43	2.89/4	0.58
1kjs	74	8.29	0.41	5.73/2	0.42
1kviA	68	2.21	0.15	2.21/1	0.15
1mkyA3	81	4.91	0.47	4.91/1	0.47
1mla_	70	3.74	0.39	2.97/2	0.55
1mn8A	84	12.38	0.44	8.69/6	0.58
1n0uA	69	5.18	0.41	4.17/2	0.51
1ne3A	56	5.29	0.3	4.66/3	0.2
1no5A	93	11.02	0.52	10.69/5	0.49
1npsA	88	2.56	0.61	2.29/3	0.63
1o2fB	77	6.91	0.52	6.91/1	0.52
1of9A	77	3.41	0.35	3.41/1	0.35
1ogwA	72	1.46	0.61	1.46/1	0.61
1orgA	118	3.06	0.50	2.51/2	0.50
1pgx_	59	3.29	0.49	3.16/2	0.59
1r69	61	2.08	0.44	1.45/4	0.45
1sfp	111	5.35	0.14	5.05/3	0.01
1shfA	59	1.48	0.54	1.48/1	0.54
1sro_	71	3.22	0.55	3.22/1	0.55
1ten_	87	2.06	0.47	2.00/3	0.42
1tfi	47	5.52	0.48	5.52/1	0.48
1thx_	108	2.98	0.49	2.32/4	0.15
1tif	59	10.02	0.50	8.19/3	0.52
1tig	88	8.59	0.59	4.36/2	0.01
1vcc	76	10.35	0.30	7.94/4	0.21
2cr7A	60	4.97	0.11	2.85/5	0.87
2f3nA	65	2.01	0.52	1.87/2	0.17
2pcy	99	4.71	0.47	4.51/2	0.24
2reb	60	10.05	0.49	5.02/2	0.52
256bA	106	3.49	0.48	3.39/2	0.55
Average	76^{m}	5.36 (2.93)	0.44	4.55 (2.68)/3^{m}	0.43

ones). For such a reduced protein set the RMSDs of the predictions are 2.93 Å and 2.68 Å, for the best among all clusters, and for the best among the top five clusters, respectively. Interestingly, often it is not the most populated cluster that is the best. The median of the distribution of best cluster among the top five is number 3. Energy minimization is required in all cases, because an average ΔRMSD (between minimized and non-minimized decoys) for the first cluster medoid was about 0.44 Å and almost the same: 0.43 Å, for the best modeled among top five medoids. Anyway, in all further analysis only minimized structures were used. Thus, by referring to a medoid of a cluster it means that this structure was energy minimized before use (no re-clustering had been made after energy minimization).

B. Protein structure quality and medoids' resolution

Figure 2 presents the correlation between model quality and quality score predicted by the neural network-based ProQ method.[58] It is noted that there is a high correlation between RMSD and TM-score, and LG-score and Max-Sub, but there is no significant correlation between RMSD or TM-score and the predicted quality of decoys. Models quality prediction was based only on a single structure, and the applied method looks only for structural defects (e.g., atom clashes). Because each medoid was energy minimized (to the nearest energy minimum), the number of such structural defects is relatively low, so evaluation consequently is poor. Thus, using scores from ProQ is not a reliable method for choosing structures for future refinement. On the basis of Figure 2 (RMSD vs TM-score) it can be seen why 4.5 Å is reasonable cut-off for trying to refine structures, since it appears that below this value there is no false-positive structures with a TM-score ≥ 0.5.

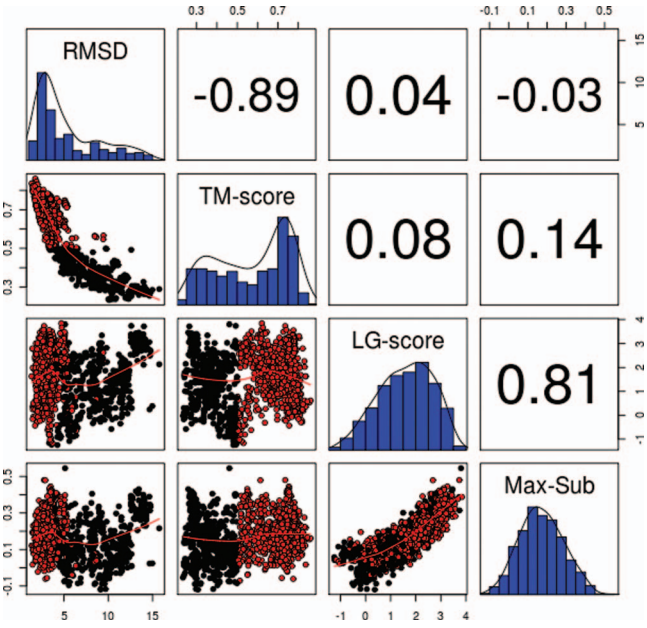

FIG. 2. Distributions and comparisons of various structure quality metrics. Upper triangle: Spearman correlation coefficients between: RMSD, TM-score, LG-score, and Max-Sub score. Diagonal: Density distributions. Lower triangle: Pairwise correlations plots between each measure. Red lines (lower triangle) are averages over all points in bins on *abscissa*. Red dots correspond to decoys for which TM-score ≥ 0.5, and black dots for these with TM-score ≤ 0.5. Plots are generated for all cluster medoids (see Sec. II).

C. Cumulative overlap of thermal motions and deformation vector

To investigate whether thermal motions are biased in the same direction as the deformation needed to obtain the native structure, the overlap between the deformation matrix of the medoid structure and the normal mode directions is calculated (see Sec. II). Analysis is performed for each trajectory in the dataset (28 cases). For each medoid, the cumulative overlap COV(k), where $k = 3, 6, 20$ between the deformation matrix and the first k normal modes is calculated. Cumulative overlaps with 5% of the slowest modes have also been calculated. One study case is presented in Figure 3. Figure presents chain A of RNA binding protein (PDB ID: 1di2) and its top medoid. For this pair of structures RMSD is equal to 2.91 Å and first three decoy's normal modes account for 0.73 fraction of deformation. The deformation origins from bad modeling of one loop (14 Gly–18 Pro) and one hairpin (27 Gly–32 Arg). It can be noticed that in both regions thermal motions (spanned by three lowest normal modes) point at a correct direction of needed deformation. For the first medoids, in the whole dataset, an average cumulative overlap is: 0.34, 0.42, 0.54, and 0.48 for $k = 3, 6, 20$, and 5% of the slowest modes, respectively. (Usually 5% of the slowest modes corresponds to $k = 11$–12.) Thus, by taking only 20 degrees of freedom, and sampling along them, we are able to derive more than 50% of the motion required for moving towards the native state. This is also nearly true for 5% of the slowest motions, which according to the JAM model should more or less describe the anharmonic motion directions.[41]

However, RMSD for the C^{α} atoms is not a fully perfect measure of how close the structure is to the native state. Not

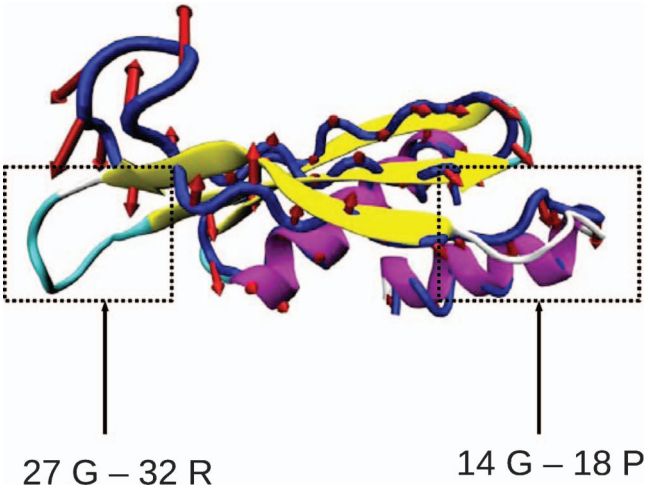

FIG. 3. An example case for an overlap between thermal motion of decoy structure and deformation from the native state. Picture presents best medoid of 1di2A protein. Native structure is in cartoon representation, colored according its secondary structure elements (purple – helix; yellow – β strand; light green – hairpin; white – loop), and decoy in represented as a blue tube. Combination of the three lowest normal modes in presented as red arrows. In boxes mark two worst modeled structure elements: loop from 14 Gly to 18 Pro and hairpin from 27 Gly to 32 Arg.

all states with low RMSD are necessarily directly (in the one step) accessible from the native state. Sometimes other more deformed structures with larger RMSDs can be more readily accessed. Because of this we decided to consider all top five medoids. These provide an even higher average cumulative overlap COV(k) equal to 0.38, 0.46, 0.63, and 0.55 for $k = 3$, 6, 20, and 5% of the slowest modes, respectively. This shows that a small increase in the number of conformations permits an even better (up to 0.63) sampling of space for moving towards the native state.

To obtain a broader perspective about the effect of a structure deformation along the normal modes, we have calculated the cumulative overlap between the deformation vector and the isotropic Gaussian deformation. In both cases: for the best medoid and the top five medoids, this value equals on average $\sim$0.21. The averaging was done for all of the proteins considered in the paper and 1000 randomly generated normal modes.

Since protein structures fluctuate, thermal motions can lead to global↔local and local↔local conformational transitions between energy minima. In the original version of the manuscript only local→global transitions were considered. Thus, it is interesting to investigate whether the cumulative overlap is the higher for transitions between local energy minima substates and the native structure (global energy minimum state) or between local energy minima substates, as asked by the reviewer. We have found that despite the fact that the cumulative overlap is high for transitions from local minima to the native state, often the highest cumulative overlap occurs for a transition between two medoids. Thus, it is interesting to investigate, whether the newly reached local energy minimum conformation is structurally closer to the native state. In our calculations all decoys for all studied proteins were used except for the worst ones (where by a definition every transition to a new conformation would be always closer to the native one). For cases where the RMSD between decoys was larger than 4.5 Å the cumulative overlap was set to zero, to avoid unlikely direct transitions between structurally distant conformations. We have found that on average, medoid conformation with the highest cumulative overlap was 0.35 Å closer to the native structure than the starting one.

The quality of the results depends on what one wishes to obtain from performing a structure refinement. If the goal is to reach experimental native structure, then following only the

TABLE II. Cumulative overlap table for medoids of the first largest clusters. The first column is the PDB name of the studied protein. The second to fourth columns show cumulative overlaps for the first 3, 6, and 20 slowest modes, respectively. The fifth column show cumulative overlap for 5% of the slowest modes, and the number of those modes. The sixth column shows Spearman correlation coefficient between positional cumulative overlap and B-factor (calculated based on decoy structure) and the seventh column shows Spearman correlation coefficient between positional cumulative overlap and root square deviation (RSD) of every amino acid. The superscript ^{m} denotes a median from the column.

Name	COV(3)	COV(6)	COV(20)	COV(5%)/NM.No.	Spearman correlation (COV, B-Factors)	Spearman correlation (COV, RSD)
1ah9_	0.21	0.37	0.54	0.39/9	0.84	0.80
1b72A	0.35	0.36	0.52	0.40/7	0.78	0.68
1cewI	0.42	0.48	0.55	0.54/15	0.74	0.88
1cqkA	0.36	0.41	0.48	0.47/14	0.73	0.81
1csp_	0.26	0.33	0.49	0.34/9	0.79	0.86
1di2A	0.65	0.76	0.79	0.76/10	0.72	0.86
1dtjA	0.24	0.39	0.45	0.41/10	0.73	0.77
1egxA	0.12	0.16	0.42	0.36/16	0.72	0.80
1fadA	0.21	0.52	0.59	0.52/13	0.75	0.71
1fo5A	0.11	0.12	0.37	0.26/12	0.76	0.60
1g1cA	0.52	0.53	0.57	0.56/14	0.69	0.63
1gyvA	0.31	0.36	0.48	0.46/17	0.65	0.85
1hbkA	0.48	0.57	0.67	0.65/13	0.87	0.78
1jnuA	0.34	0.43	0.63	0.60/15	0.78	0.69
1kviA	0.15	0.16	0.38	0.19/9	0.80	0.56
1mla_	0.27	0.28	0.38	0.32/10	0.74	0.61
1npsA	0.37	0.40	0.59	0.54/12	0.77	0.77
1of9A	0.42	0.55	0.64	0.58/11	0.71	0.69
1ogwA	0.48	0.55	0.65	0.56/10	0.81	0.73
1orgA	0.40	0.52	0.57	0.56/17	0.78	0.67
1pgx_	0.31	0.44	0.51	0.45/8	0.66	0.73
1r69	0.39	0.40	0.48	0.40/8	0.73	0.60
1shfA	0.21	0.39	0.64	0.44/8	0.80	0.75
1sro_	0.36	0.42	0.52	0.45/10	0.82	0.74
1ten_	0.45	0.46	0.57	0.54/12	0.73	0.82
1thx_	0.21	0.24	0.45	0.44/15	0.82	0.77
2f3nA	0.54	0.57	0.67	0.59/9	0.78	0.71
256bA	0.40	0.45	0.59	0.57/15	0.83	0.82
Average	0.34	0.42	0.54	$0.48/11.5^{m}$	0.76	0.74

TABLE III. Cumulative overlap for the single best medoid from the first five clusters. The first column shows the PDB name of the studied protein. The second column shows the index number of the best cluster (according to COV(20)) and its RMSD. The third to fifth columns show the cumulative overlaps for 3, 6, and 20 slowest modes, respectively. The sixth column shows the cumulative overlap for 5% of the slowest modes, and also the number of these modes. The seventh column shows Spearman correlation coefficient between positional cumulative overlap and B-factor (calculated based on decoy structure) and the eighth column shows Spearman correlation coefficient between positional cumulative overlap and root square deviation (RSD) of every amino acid. The superscript ^{m} denotes a median from the column.

Name	Top 5 #	COV(3) Top 5	COV(6) Top 5	COV(20) Top 5	COV(5%)/NM.No.	Spearman correlation (COV, B-factors)	Spearman correlation (COV, RSD)
1ah9_	1/3.42	0.21	0.37	0.54	0.39/9	0.84	0.80
1b72A	3/4.56	0.50	0.50	0.66	0.58/7	0.69	0.68
1cewI	4/4.11	0.40	0.48	0.55	0.51/15	0.75	0.85
1cqkA	2/2.64	0.39	0.42	0.62	0.58/14	0.82	0.82
1csp_	5/2.58	0.26	0.37	0.64	0.52/9	0.79	0.74
1di2A	2/3.11	0.73	0.75	0.80	0.78/10	0.77	0.80
1dtjA	5/2.65	0.41	0.54	0.69	0.57/10	0.80	0.79
1egxA	3/2.03	0.09	0.20	0.51	0.39/16	0.73	0.83
1fadA	5/3.69	0.35	0.59	0.63	0.61/13	0.84	0.60
1fo5A	3/3.84	0.09	0.14	0.47	0.25/12	0.84	0.61
1g1cA	5/3.39	0.43	0.50	0.59	0.56/14	0.77	0.85
1gyvA	2/3.60	0.27	0.35	0.50	0.45/17	0.69	0.89
1hbkA	1/3.95	0.48	0.57	0.67	0.65/13	0.87	0.78
1jnuA	1/3.27	0.34	0.43	0.63	0.60/15	0.78	0.69
1kviA	2/2.49	0.45	0.46	0.61	0.48/9	0.71	0.70
1mla_	2/2.96	0.40	0.42	0.60	0.45/10	0.75	0.67
1npsA	1/2.56	0.37	0.40	0.59	0.54/12	0.77	0.77
1of9A	3/3.74	0.44	0.59	0.70	0.63/11	0.75	0.66
1ogwA	1/1.46	0.48	0.55	0.65	0.56/10	0.81	0.73
1orgA	3/2.94	0.28	0.53	0.62	0.61/17	0.79	0.70
1pgx_	2/3.16	0.44	0.51	0.59	0.55/8	0.66	0.80
1r69	5/2.23	0.48	0.57	0.62	0.59/8	0.80	0.71
1shfA	3/1.58	0.36	0.39	0.76	0.51/8	0.75	0.78
1sro_	4/3.92	0.20	0.42	0.64	0.53/10	0.81	0.84
1ten_	2/2.72	0.41	0.53	0.67	0.65/12	0.85	0.81
1thx_	4/2.32	0.31	0.33	0.57	0.54/15	0.77	0.81
2f3nA	4/2.16	0.64	0.66	0.78	0.72/9	0.87	0.69
256bA	3/3.53	0.41	0.46	0.63	0.58/15	0.86	0.86
Average	3^{m}/3.02	0.38	0.46	0.63	0.55/11.5^{m}	0.78	0.76

slowest normal modes, and using only a single medoid would not readily yield a significant structural refinement. Nonetheless, by selecting only a few medoids structures and by deforming these along only a few normal modes it is possible to obtain a significant enrichment of the near-native conformational space for structural refinement.

Other ways to possibly enhance the sampling might include some deviations from the normal mode directions. Such putative sampling would be steered overall by the normal modes, but some fluctuations around these directions are allowed. However, the demonstration above that 63% of the samples move toward the native structure is a relatively high performance and adding (computationally costly) random motions would be unlikely to improve these results.

The intended use of the present scheme is to generate an ensemble of structures and to evaluate them (depending upon the resolution of the generated decoys) using either a coarse-grained or an atomic force-field, and follow this up with energy minimization. It has to be mentioned that ANM samples structures based on the harmonic approximation, with the starting structure having the lowest energy. This is certainly not the case if the applied deformations are large, and a structure traverses energy barriers between two energy minima basins. This conclusion can be reached with the JAM model although harmonic modes are not exactly the same as intersubstate modes, but they actually vibrate in similar directions as the anharmonic intersubstates.[41] Further analysis of the positional cumulative overlap and thermal mobility and RMSD for each position is consistent with a hierarchical view of conformational substates.[61] Positional cumulative overlaps describe how much each amino acid, here represented only by C^{α}, contributes to COV(k) (see Tables II and III). Interestingly, it was found that for both cases: of the first medoid and the top five cluster medoids, correlations between positional overlap in mobility or deviation from the native state is quite significant and equals: 0.76, 0.74 and 0.78, 0.76 for first and the top five medoids, respectively. This means that amino acids that are more mobile or deviate more from the native structure, move significantly in directions toward native state. This is also supported by a hierarchical model of conformational substates since jumps between two energy basins do not involve changes equally throughout the whole structure, but only changes in few parts of the structure – such as closing and opening hinges in many enzymes. This finding

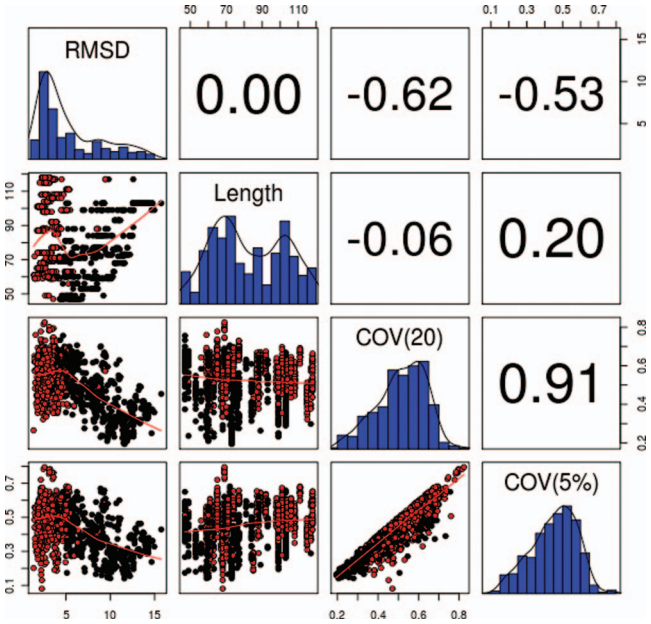

FIG. 4. Distributions and comparisons of RMSD, sizes and cumulative overlaps. Upper triangle: Spearman correlation coefficients for: RMSD, Length, COV(20), and COV(5%). Diagonal: The density distributions for each quantity. Lower triangle: Pairwise correlations plots between each measure. Red lines (lower triangle) are averages over all points in bins on *abscissa*. Red dots correspond to decoys for which TM-score ≥ 0.5, and black dots for cases with TM-score ≤ 0.5. Plots are generated for all cluster medoids (see Sec. II).

can be also useful for developing structure refinement algorithms,[16,62] where only some parts of the chain are deformed and the core domain residues are kept fixed.

D. Quality of the model and cumulative overlap

The lack of the correlation between RMSD and the sizes of the modeled structures in Figure 4 suggests that the predictive power of the I-TASSER force-field and clustering method does not actually depend much on the length of protein at least in the range of analyzed cases: 50–120 amino acids. Thus it may suggest that the results are related to generic properties of protein substates and that these are independent of size. The same is true regarding the correlation between protein length and COV(20) or COV(5%). Because usually 5% of modes include fewer than 20 modes the deviations away from linearity in Figure 4 for COV(20) vs COV(5%) is easily understood. But in summary COV(20) is only slightly different from COV(5%). It can be explained by the fact, that 5% of the lowest modes account for the majority of the system's variance. Another interesting fact is that there is a fairly high correlation between both COV(20) or COV(5%) and RMSD. The correlation is caused by a tail of results for decoys modeled with resolution from 4 to 5 Å (i.e., when TM-score is lower than 0.5) to 15 Å. These models definitely are not suitable for refinement using normal modes. However, for models with resolution below 4 Å, the correlation is near zero. Some of decoys have cumulative overlap as high as 0.8 but some of them as low as 0.1. The explanation of this fact comes from hierarchical picture of conformational substates. It means that some structures must traverse other substates' basin before reaching the native state, so they are not directly connected to the native basin through a simple fluctuation type pathway; i.e., the pathway on the energy surface has more than one saddle point.

IV. CONCLUSIONS

Our results suggest some directions for further development of refinement algorithms. Such an algorithm could be for example MD or MC steered by normal modes. In these approaches one would deform a given decoy, or set of decoys, along a few slowest normal modes. The magnitude of such deformation could be either uniform along the chain or dependent on mobility of each residue separately, or could even depend on other parameters such as the accessible surface area (ASA), secondary structure, or others. Beside, as can be seen from Eq. (7) cumulative overlap is a measure between directions of vectors. The normal modes are orthogonal, so combining them requires specification of phase angles between the pairs of modes. These phase angles could in principle be determined from simulations by MD or MC. To improve the accuracy of refinement, the most promising approach would be to pursue an ensemble of different structures and try to refine them, and in the end pick up the best model among the all refined.

Another significant problem encountered in protein refinement is imperfect force field used to compute energies of protein models. Recently, we have shown that a random noise in the force field can prevent protein refinement if the level of noise is too high.[63] Thus, by reducing the number of degrees of freedom, we diminish the influence of errors in force field on the structure refinement.

We have demonstrated that there are significant gains that can be obtained application of the sampling of conformations with the elastic network models. The most critical consideration for successful application of this approach is to have a starting conformation that is reasonably close, say within 4 Å of the native structure.

Since protein motions are damped by both intermolecular interactions with solvent and intramolecularly, it is possible to represent such effects by extending our approach with an overdamped Langevin equation.

ACKNOWLEDGMENTS

We thank Michael T. Zimmermann at ISU for helpful comments on this manuscript. We would like to acknowledge support from National Institutes of Health Grant Nos. R01GM072014, R01GM073095, R01GM081680, and R01GM081680-S1, and from National Science Foundation Grant No. NSF MCB 1021785. A. Kolinski acknowledges support from Foundation for Polish Science, Grant No. TEAM/2011-7/6 cofinanced by the European Regional Development Fund operated within the Innovative Economy Operational Program. Computational part of this work was carried out using the computer cluster at the Computing Centre of the Faculty of Chemistry, University of Warsaw.

[1] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard, Proteins **53**, 334 (2003).
[2] J. Skolnick and M. Brylinski, Briefings Bioinf. **10**, 378 (2009).
[3] M. Brylinski and J. Skolnick, PLOS Comput. Biol. **5**(6), e1000405 (2009).
[4] I. Bahar, T. R. Lezon, L. W. Yang, and E. Eyal, Annu. Rev. Biophys. **39**, 23 (2010).
[5] R. F. Service, Science **319**, 1610 (2008).
[6] A. Roy, A. Kucukural, and Y. Zhang, Nat. Protoc. **5**, 725 (2010).
[7] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[8] L. Yang, G. Song, and R. L. Jernigan, Biophys. J. **93**, 920 (2007).
[9] I. Bahar, T. R. Lezon, A. Bakan, and I. H. Shrivastava, Chem. Rev. **110**, 1463 (2010).
[10] I. Bahar, C. Chennubhotla, and D. Tobi, Curr. Opin. Struct. Biol. **17**, 633 (2007).
[11] C. J. Tsai, B. Ma, and R. Nussinov, Proc. Natl. Acad. Sci. U.S.A. **96**, 9970 (1999).
[12] I. T. Jolliffe, *Principal Component Analysis* (Springer, New York, 2002).
[13] O. Marques and Y. H. Sanejouand, Proteins **23**, 557 (1995).
[14] D. Perahia and L. Mouawad, Comput. Chem. **19**, 241 (1995).
[15] I. Bahar, B. Erman, T. Haliloglu, and R. L. Jernigan, Biochemistry **36**, 13512 (1997).
[16] F. Tama and T.-H. Sanejouand, Protein Eng. **14**, 1 (2001).
[17] M. K. Kim, R. L. Jernigan, and G. S. Chirikjian, Biophys. J. **83**, 1620 (2002).
[18] M. K. Kim, R. L. Jernigan, and G. S. Chirikjian, J. Mol. Graphics Modell. **21**, 151 (2002).
[19] M. K. Kim, R. L. Jernigan, and G. S. Chirikjian, Biophys. J. **89**, 43 (2005).
[20] C. Xu, D. Tobi, and I. Bahar, J. Mol. Biol. **333**, 153 (2003).
[21] L. Orellana, M. Rueda, C. Ferrer-Costa, J. R. Lopez-Blanco, P. Chacon, and M. Orozco, J. Chem. Theory Comput. **63**, 2910 (2010).
[22] T. Z. Sen, Y. Feng, J. Garcia, A. Kloczkowski, and R. L. Jernigan, J. Chem. Theory Comput. **2**, 696 (2006).
[23] A. Yan, Y. Wang, A. Kloczkowski, and R. L. Jernigan, J. Chem. Theory Comput. **4**, 1757 (2008).
[24] O. Kurkcuoglu, P. Doruker, T. Z. Sen, A. Kloczkowski, and R. L. Jernigan, Phys. Biol. **5**, 046005 (2008).
[25] K. Steczkiewicz, M. Kurcinski, M. T. Zimmermann, B. Lewis, D. Dobbs, A. Kloczkowski, R. L. Jernigan, A. Kolinski, and K. Ginalski, Proc. Natl. Acad. Sci. U.S.A. **108**, 9443 (2011).
[26] M. T. Zimmermann, A. Skliros, A. Kloczkowski, and R. L. Jernigan, Immunome Res. **7**, 5 (2011).
[27] M. T. Zimmermann, A. Kloczkowski, and R. L. Jernigan, BMC Bioinf. **12**, 264 (2011).
[28] M. T. Zimmermann, S. P. Leelananda, P. Gniewek, Y. P. Feng, R. L. Jernigan, and A. Kloczkowski, J. Struct. Funct. Genomics **12**, 137 (2011).
[29] A. Skliros, M. T. Zimmermann, D. Chakraborty, S. Saraswathi, A. R. Katebi, S. Leelananda, A. Kloczkowski, and R. L. Jernigan, Phys. Biol. **9**, 014001 (2012).
[30] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern, Nature (London) **438**, 117 (2005).
[31] W. L. Hubbell, D. S. Cafiso, and C. Altenbach, Nat. Struct. Biol. **7**, 735 (2000).
[32] G. E. Fanucci and D. S. Cafiso, Curr. Opin. Struct. Biol. **16**, 644 (2006).
[33] O. B. Ptitsyn, Adv. Protein Chem. **47**, 83 (1995).
[34] Z. Luthey-Schulten and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[35] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, Annu. Rev. Biophys. **37**, 289 (2008).
[36] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[37] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[38] E. J. Sorin and V. S. Pande, Biophys. J. **88**, 2472 (2005).
[39] G. Chopra, C. M. Summa, and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **105**, 20239 (2008).
[40] A. W. Stumpff-Kane and M. Feig, Proteins **63**, 155 (2006).
[41] A. Kitao, S. Hayward, and N. Go, Proteins **33**, 496 (1998).
[42] J. Deak, H.-L. Chiu, C. M. Lewis, and R. J. Dwayne-Miller, Phys. Chem. B **102**, 6621 (1998).
[43] J. Ma, Structure **13**, 373 (2005).
[44] R. Elber and M. Karplus, Chem. Phys. Lett. **139**, 375 (1987).
[45] J. I. Sulkowska, A. Kloczkowski, T. Z. Sen, M. Cieplak, and R. L. Jernigan, Proteins **71**, 45 (2008).
[46] Y. Feng, L. Yang, A. Kloczkowski, and R. L. Jernigan, Proteins **77**, 551 (2009).
[47] Y. Zhang, BMC Bioinf. **9**, 40 (2008).
[48] A. Kolinski, Acta Biochim. Pol. **51**, 349 (2004).
[49] P. Bradley, D. Chivian, J. Meiler, K. M. Misura, C. A. Rohl, W. R. Schief, W. J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C. E. Strauss, and D. Baker, Proteins **53**, 457 (2003).
[50] J. L. MacCallum, L. Hua, M. J. Schneiders, V. S. Pande, M. P. Jacobson, and K. A. Dill, Proteins **77**(S9), 66 (2009).
[51] J. L. MacCallum, A. Perez, M. J. Schnieders, L. Hua, M. P. Jacobson, and K. A. Dill, Proteins **79**(S10), 74 (2011).
[52] D. Shortle, K. T. Simons, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **95**, 11158 (1998).
[53] D. Gront and A. Kolinski, Bioinformatics **21**, 3179 (2005).
[54] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).
[55] J. Zhang and Y. Zhang, PLoS One **5**, e15386 (2010).
[56] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996).
[57] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[58] B. Wallner and A. Elofsson, Protein Sci. **12**, 1073 (2003).
[59] D. T. Jones, J. Mol. Biol. **292**, 195 (1999).
[60] J. Xu and Y. Zhang, Bioinformatics **26**, 889 (2010).
[61] T. Noguti and N. Go, Proteins **5**, 132 (1989).
[62] A. W. Stumpff-Kane, K. Maksimiak, S. L. Lee, and M. Feig, Proteins **70**, 1345 (2005).
[63] P. Gniewek, A. Kolinski, R. L. Jernigan, and A. Kloczkowski, Proteins **80**, 335 (2012).
[64] Y. Zhang and J. Skolnick, Proteins **57**, 702 (2004).
[65] E. Eyal, L.-W. Yang, and I. Bahar, Bioinformatics **22**, 2619 (2006).
[66] M. Delarue and P. Dumas, Proc. Natl. Acad. Sci. U.S.A. **101**, 6957 (2004).
[67] F. Pontiggia, G. Colombo, C. Micheletti, and H. Orland, Phys. Rev. Lett. **98**, 048102 (2007).
[68] F. Pontiggia, A. Zen, and C. Micheletti, Biophys. J. **95**, 5901 (2008).