THE JOURNAL OF CHEMICAL PHYSICS **133**, 155101 (2010)

Sequence composition and environment effects on residue fluctuations in protein structures

Anatoly M. Ruvinsky[1,a)] and Ilya A. Vakser[1,2]
[1]*Center for Bioinformatics, The University of Kansas, Lawrence, Kansas 66047, USA*
[2]*Department of Molecular Biosciences, The University of Kansas, Lawrence, Kansas 66047, USA*

(Received 13 May 2010; accepted 16 September 2010; published online 15 October 2010)

Structure fluctuations in proteins affect a broad range of cell phenomena, including stability of proteins and their fragments, allosteric transitions, and energy transfer. This study presents a statistical-thermodynamic analysis of relationship between the sequence composition and the distribution of residue fluctuations in protein-protein complexes. A one-node-per-residue elastic network model accounting for the nonhomogeneous protein mass distribution and the interatomic interactions through the renormalized inter-residue potential is developed. Two factors, a protein mass distribution and a residue environment, were found to determine the scale of residue fluctuations. Surface residues undergo larger fluctuations than core residues in agreement with experimental observations. Ranking residues over the normalized scale of fluctuations yields a distinct classification of amino acids into three groups: (i) highly fluctuating-Gly, Ala, Ser, Pro, and Asp, (ii) moderately fluctuating-Thr, Asn, Gln, Lys, Glu, Arg, Val, and Cys, and (iii) weakly fluctuating-Ile, Leu, Met, Phe, Tyr, Trp, and His. The structural instability in proteins possibly relates to the high content of the highly fluctuating residues and a deficiency of the weakly fluctuating residues in irregular secondary structure elements (loops), chameleon sequences, and disordered proteins. Strong correlation between residue fluctuations and the sequence composition of protein loops supports this hypothesis. Comparing fluctuations of binding site residues (interface residues) with other surface residues shows that, on average, the interface is more rigid than the rest of the protein surface and Gly, Ala, Ser, Cys, Leu, and Trp have a propensity to form more stable docking patches on the interface. The findings have broad implications for understanding mechanisms of protein association and stability of protein structures. © *2010 American Institute of Physics*. [doi:10.1063/1.3498743]

I. INTRODUCTION

A remarkable difference between sequence compositions of regular and irregular secondary structure elements of proteins has been attracting considerable attention for more than 30 years.[1–6] Amino-acid composition profiles revealed that the irregular regions (protein loops) are enriched in Gly, Pro, Ser, and Asp. The regular regions (α-helices and β-strands) contain less of these amino acids. Helices are enriched in Leu, Ala, Glu, and Gln, and β-strands are enriched in Val, Ile, Phe, and Tyr. Amino-acid composition of protein interfaces has been analyzed.[7–10] Despite the extensive use of the statistics in almost all aspects of protein modeling (e.g., in computational algorithms for the secondary structure assignments (see Ref. 11 for the review), in knowledge-based approaches to prediction of protein structures[12,13] receptor-ligand docking[14–16]), the understanding of mechanisms underlying and amino-acid propensities is still incomplete and poses a challenge for researchers in physics and biology. Recent discoveries of chameleon sequences, that undergo helix-sheet transitions,[17–21] and intrinsically disordered proteins or fragments, that undergo order-disorder transitions,[22–24] have added interest to the problem. One way to tackle this puzzle is to study the distribution, the scale, and features of structural and thermal fluctuations in proteins.

Protein functionality, encoded into the sequence, is based on a dual ability of proteins to sustain and change their structures.[25] The relationship has different degrees of sensitivity to the location and the scale of changes in protein structures (e.g., CH_3 group rotations, conversions of side-chain rotamers, cis-trans isomerization of proline, or domain shifts). Last 10 years demonstrated increasing popularity of low-resolution or coarse-grained models in conjunction with harmonic potentials, called elastic network models (ENM), for deciphering and modeling various large-scale structural changes (e.g., allosteric changes in protein structures,[26–29] structural changes on transition pathways,[26,30–36] and global conformational changes upon protein-protein binding[37–39]). Other applications of these models include the analysis of Debye–Waller factors of C_α atoms,[29,40–45] protein-protein[38,46] and protein-ligand[47] docking, x-ray crystallographic refinement,[48] and structural variations in ensembles of NMR structures.[49,50]

Two types of ENMs are widely used: homogeneous and nonhomogeneous models. A homogeneous ENM is a network of nodes represented by C_α atoms and connected by Hooke springs if the distance between nodes is less than a cutoff radius.[26,29,31,35–37,39,41,42] All network nodes are as-

[a)]Electronic mail: ruvinsky@ku.edu.

signed an equal mass that smoothes protein mass density. The homogeneous ENM has two parameters only, the cutoff radius and the spring force constant. Nonhomogeneous ENMs introduce structural and interaction inhomogeneity by assigning residue masses to the network nodes represented by C_α atoms[34,45] or by assigning distance- or residue type-dependent force constants to interacting nodes.[28,30,32,34,40,44–46,51] The effect of protein sequence variations on the spring force constants has been considered recently.[27] Double-well ENMs are used to model large-scale conformational transition pathways.[30,31,36,52] Merging residues into rigid blocks is used to consider properties of large macromolecules within ENM of a lower resolution.[26,35,43,53] Less "extreme" coarse graining keeps three translational degrees of freedom of C_α-based nodes and degrees of freedom of bond angles and dihedrals (see Refs. 52, 54, and 55).

In the context of nonhomogeneous ENMs, we present a novel method to account for the protein mass distribution and interatomic contacts within the coarse-grained model. We move network nodes from C_α atoms to the centers of mass of protein residues to bring in the effects of side chains into the model. We derive a modified Tirion-like potential[56] to bring in structural details of the atomic level and put forward a statistical-thermodynamic formalism to calculate residue fluctuations in a set of protein complexes.[57] We show that the scale of residue fluctuations increases from the core to the surface of a protein in agreement with the experimental data.[58–60] We suggest a classification of protein residues based on the normalized scale of fluctuations and discuss how the scale of fluctuations correlates with amino acid propensities in the secondary structure elements, chameleon sequences, and disordered fragments. Fluctuations of binding site residues (interface residues) are compared with other surface residues. The tendency of some residues to form more stable docking patches on the interface is discussed as well as the role of loops at early stages of protein thermal denaturation.

II. MATERIALS AND METHODS

A modified nonhomogeneous ENM is used in calculations. Network nodes are placed in the centers of mass of protein residues and residue masses are assigned to the corresponding network nodes. The following is a description of a formalism to consistently transform the interatomic protein energy landscape into the inter-residue landscape. As a result, we obtain a modified inter-residue harmonic potential with a spring force constant proportional to the number of interatomic contacts between residues [see Eq. (3) below].

The interaction energy between protein residues i and k is

$$U_{ik}(\vec{R}_i-\vec{R}_k)=\sum_{\alpha,\beta}U_{\alpha\beta}(\vec{R}_i+\vec{u}^i_\alpha-\vec{R}_k-\vec{u}^k_\beta), \tag{1}$$

where $\vec{R}_{i,k}$ are radius vectors of the centers of mass of residues i and k and $\vec{u}^{i,k}_{\alpha,\beta}$ are the radius vectors of atoms α and β relative to the centers of mass of the residues i and k accordingly. The sum in Eq. (1) runs over all pairs of atoms separated by a distance less than the interaction cutoff. Introducing a residue-residue potential, one can rewrite Eq. (1) as $U_{ik}(\vec{R}_i-\vec{R}_k)=N_{ik}V(\vec{R}_i-\vec{R}_k)$, where N_{ik} is a number of interatomic interactions between residues i and k and V is averaged interatomic potential. Assuming that inter-residue interactions are in equilibrium in the native protein and using a Lennard-Jones form of the inter-residue potential, we can expand $V(\vec{R}_i-\vec{R}_k)$ in Taylor series of deviations $R_{ik}-R^0_{ik}$ of the inter-residue distance $R_{ik}=|\vec{R}_i-\vec{R}_k|$ from its equilibrium R^0_{ij}. Expanding to the second order in $R_{ij}-R^0_{ij}$ yields

$$U_{ik}(\vec{R}_i-\vec{R}_k)=-N_{ik}\varepsilon+36N_{ik}\varepsilon\left(\frac{R_{ik}-R^0_{ik}}{R^0_{ik}}\right)^2, \tag{2}$$

where ε is the depth of the Lennard-Jones potential. Equation (2) shows that inter-residue interactions are proportional to the number of interatomic interactions and decrease with the increase of the inter-residue distance as $1/(R^0_{ik})^2$.

Since $\vec{R}_{i,k}=\vec{R}^0_{i,k}+\vec{r}_{i,k}$, we obtain

$$U_{ik}(\vec{r}_i-\vec{r}_k,\vec{R}^0_{ik})=-\varepsilon N_{ik}+36\varepsilon\frac{N_{ik}}{(R^0_{ik})^2}\left(\frac{\vec{R}^0_{ik}}{R^0_{ik}}(\vec{r}_i-\vec{r}_k)\right)^2, \tag{3}$$

where $\vec{r}_{i,k}$ are the deviations of the residue centers of mass from its equilibrium position. The main difference between Eq. (3) and Tirion-like potentials[56] used in nonhomogeneous ENMs is the factor N_{ik} which introduces the distribution of interatomic interactions into the coarse-grained model. In other words, the change of the protein model resolution from the atomic to the residue level results in the appearance of this factor in the inter-residue potential. The important role of the factor N_{ik} is supported by the local density model[61] that was shown to be efficient in predicting atomic fluctuations and B-factors. The nonbonded potential energy of an atom in the local density model is proportional to the number of interactions with noncovalent neighbor atoms. Relationship of the optimized spring force constants to the average number of the nearest C_α atoms was shown to be important for the Gaussian ENM.[51]

The protein Lagrangian,

$$\mathcal{L}=\sum_{i,k=1}^{N}\frac{m_i}{2}\left(\frac{d\vec{r}_i}{dt}\right)^2-U_{ik}(\vec{r}_i-\vec{r}_k,\vec{R}^0_{ik}), \tag{4}$$

derives the following $3N$ equations of motions:

$$m_i\ddot{\vec{r}}_i=-\sum_{k=1}^{N}C_{ik}(\vec{\alpha}_{ik}(\vec{r}_i-\vec{r}_k))\vec{\alpha}_{ik}, \tag{5}$$

where m_i is the mass of the residue i, $\vec{\alpha}_{ik}=\vec{R}^0_{ik}/R^0_{ik}$, $C_{ik}=72\varepsilon N_{ik}/(R^0_{ik})^2$, and N is the number of protein residues. As usual, following classical methodology (see Refs. 62–64 or elsewhere), we seek an oscillatory solution of the form $\vec{r}_k=\vec{A}_k\exp(i\omega t)$, where A_k are some amplitude factors to be determined. The substitution of the trial solution into the equations of motions leads to the eigenvalue problem $(\mathbf{H}-\omega^2\mathbf{I})\mathbf{A}=0$, where $\mathbf{A}=\{A^x_1,A^y_1,A^z_1,A^x_2,A^y_2,\ldots\}$ is a $3N$ column vector of the amplitude factors, $\mathbf{I}$ is a $3N\times3N$ unit matrix, and $\mathbf{H}$ is a $3N\times3N$ matrix composed of 3×3 superelements,

$$H_{ik}(i \neq k) = \begin{bmatrix} h_{ik}^{xx} & h_{ik}^{yx} & h_{ik}^{zx} \\ h_{ik}^{xy} & h_{ik}^{yy} & h_{ik}^{zy} \\ h_{ik}^{xz} & h_{ik}^{yz} & h_{ik}^{zz} \end{bmatrix}, \quad H_{ii} = -\sum_k{}' H_{ik}, \tag{6}$$

where $h_{ik}^{ab} = -C_{ik}\alpha_{ik}^a \alpha_{ik}^b / m_i$ and the upper indexes a,b stand for x,y,z projections of the vector $\vec{\alpha}_{ik}$.

The prime in sums over k in Eqs. (6) means that a term $i=k$ is not accounted for. We use our program to find protein eigenfrequencies $\{\omega\}$ and normalized eigenvectors. The kth oscillation can be written in the form

$$x_k = \sum_{i=1}^{3N-6} G_{ki} c_i \exp(\omega_i t) = \sum_{i=1}^{3N-6} G_{ki}\Theta_i, \tag{7}$$

where $\Theta_i = Re[c_i \exp(\omega_i t)]$ is the so-called normal coordinate, Re stands for "real part of," c_i is a constant determined by initial conditions, and columns of the matrix $\mathbf{G}$ are the normalized eigenvectors. The normal modes are described by

$$\mathcal{H} = \sum_{i=1}^{3N-6} \frac{M_i}{2}(\dot{\Theta}_i^2 + \omega_i^2\Theta_i^2), \tag{8}$$

where $M_i = \Sigma_{k=1}^{3N-6} m_k G_{ki}^2$ is the effective mass of the ith normal mode.[62] Note that for a homogeneous ENM, m_i is a constant equal to some parameter m and, therefore, all modes will have equal effective masses: $M_i = \Sigma_{k=1}^{3N-6} m G_{ki}^2 = m$.

The mean-square fluctuation of the kth residue along the coordinate axis x is $\langle x_{k_x}^2 \rangle = \Sigma_{ij} G_{k_x i} G_{k_x j} \langle \Theta_i \Theta_j \rangle$,[65] where the angular brackets denote a Boltzmann average with Hamiltonian (8) over the normal modes and $k_{x,y,z}$ are the numbers of degrees of freedom associated with the residue center of mass oscillations along the coordinate axes x,y,z. Boltzmann averaging of pair products $\langle \Theta_i \Theta_j \rangle$ of normal coordinates yields $\langle \Theta_i \Theta_j \rangle = \delta_{ij} T k_B / (M_i \omega_i^2)$, where T is the temperature, k_B is the Boltzmann constant, and δ_{ij} is the Kronecker delta ($\delta_{ij}=1$ if $i=j$ and $\delta_{ij}=0$ if $i \neq j$). The total mean-square fluctuation of the kth residue has the form

$$\langle (\vec{r}_k)^2 \rangle = T k_B \sum_{i=1}^{3N-6} \frac{G_{k_x i}^2 + G_{k_y i}^2 + G_{k_z i}^2}{M_i \omega_i^2}. \tag{9}$$

It is important to note that the residue fluctuation, derived in Eq. (9), shows nonlocal dependence on the mass distribution in a protein. This effect totally disappears in the framework of a homogeneous ENM.

Removing the effect of the parameter ε on the residue fluctuations and normalizing them, we introduce a mobility ratio (MR) of the kth residue in the form

$$\mathcal{R}_k = \frac{\langle (\vec{r}_k)^2 \rangle}{(\vec{r}^2)_{\mathrm{av}}}, \tag{10}$$

where $(\vec{r})_{\mathrm{av}}^2 = \Sigma_{k=1}^N \langle (\vec{r}_k)^2 \rangle / N$ is the averaged mean-square fluctuation in a protein.

We computed the mobility ratios for each of the protein residues in 184 proteins from the 92 nonobligate protein-protein complexes selected from a docking benchmark set.[57] For each of the proteins, MRs were grouped in 20 groups according to names of standard amino acids and 20 average MRs were computed. The obtained values were averaged over the set of 184 protein structures. Figures 2–4 show mean MRs and standard deviations.

III. RESULTS

A. Interaction cutoff

C_α-based ENMs are commonly used for predicting B-factors of C_α atoms. Often, an interaction cutoff in C_α-based ENMs is chosen in a such way that it maximizes the correlation between the B-factors of C_α atoms and the predicted fluctuations of C_α-based nodes. In our case, however, it would seem that one should not expect high correlation between the B-factors of C_α atoms and fluctuations of the residue centers of mass. Indeed, the mean-square fluctuation of the center of mass of the kth residue is

$$\begin{aligned} \langle (\vec{r}_k)^2 \rangle &= \left\langle \left(\sum_\alpha \frac{m_\alpha}{M_k} \vec{u}_\alpha^k \right)^2 \right\rangle \\ &= \sum_{\alpha=1}^{N_\alpha} \frac{m_\alpha^2}{M_k^2} \langle (\vec{u}_\alpha^k)^2 \rangle + \sum_{\alpha,\beta} \frac{2m_\alpha m_\beta}{M_k^2} \langle \vec{u}_\alpha^k \vec{u}_\beta^k \rangle \\ &= \frac{3}{8\pi^2} \sum_{\alpha=1}^{N_\alpha} \frac{m_\alpha^2}{M_k^2} B_\alpha + \sum_{\alpha,\beta} \frac{2m_\alpha m_\beta}{M_k^2} \langle \vec{u}_\alpha^k \vec{u}_\beta^k \rangle, \end{aligned} \tag{11}$$

where m_α is the atomic mass and B_α is the B-factor of the atom α. Equation (11) shows that the fluctuation of the residue center of mass depends on the B-factors of all atoms of the residue and the pair correlations of atom motions. Figure 1 illustrates the typical behavior of the correlation between the B-factors of C_α atoms and the fluctuations of the residue centers of mass [Eq. (10)] as a function of the cutoff for three proteins (189l, 1a3h, 1cwy). The proteins are of significantly different sizes (164, 303, and 500 residues accordingly). The correlation coefficients for a subset of other structures from the docking benchmark set are given in Supplementary Table I.[66] It is interesting to note that despite the complex relation between $\langle (\vec{r}_k)^2 \rangle$ and B_α [Eq. (11)], for all the proteins the correlation remains high until 10 Å and then decreases. The values of the correlation coefficient in the vicinity of the

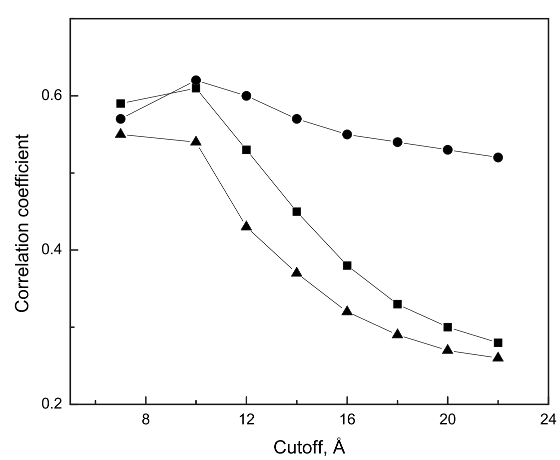

FIG. 1. Correlation coefficients between B-factors of C_α atoms and fluctuations of the residue centers of mass as a function of the cutoff for 1a3h (circles), 1cwy (triangles), and 189l (squares).

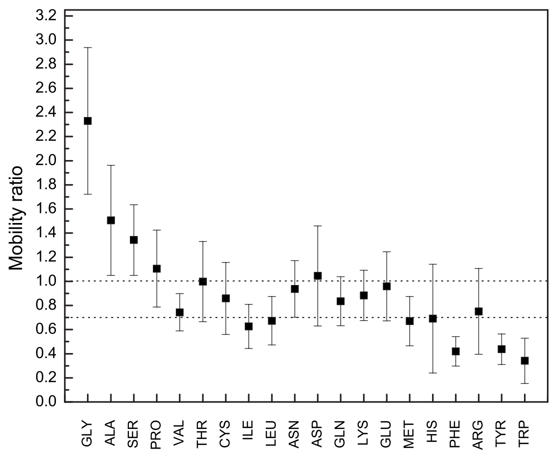

FIG. 2. The mobility ratios of protein residues calculated using Eq. (10) arranged in the order of increasing mass. The error bars show the standard deviations.

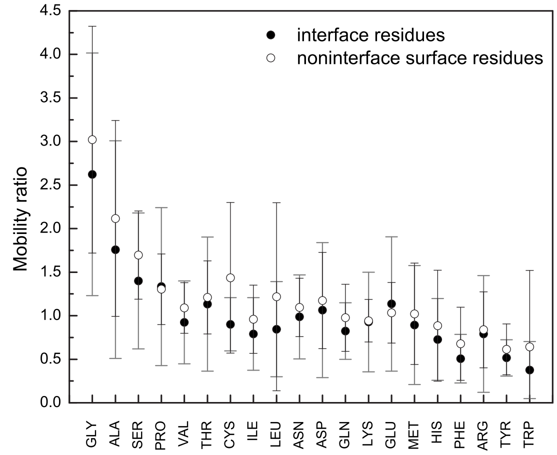

FIG. 4. The mobility ratios of interface and noninterface surface residues. The error bars show the standard deviations.

maximum are similar to the ones reported by others [e.g., see a comparative study of three common ENMs (Ref. 67)]. Riccardi *et al.*[67] showed that the average correlation between B-factors and the calculated fluctuations using the C_α ENMs vary between 0.39 and 0.63, depending on the way to account for the crystal environment. We obtained the average correlation coefficient of 0.55 [see Supplementary Table I (Ref. 66)]. Higher correlations can be achieved by using ensembles of NMR structures.[49] Therefore, the 10 Å cutoff was chosen for prediction of the mobility ratios.

Our choice of the cutoff is supported by protein studies at both all-atom and coarse-grained resolutions. Indeed, since the 10 Å cutoff is often used in all-atom molecular dynamics of proteins (e.g., Refs. 68–70) and corresponds to the last peak of the pair distribution function in proteins,[71] intuitively one could expect structural importance of the interactions up to this cutoff for low-resolution models as well. The reason is that the change of the resolution should not change the scale of interactions due to the conservation of energy. The 10 Å cutoff was recommended for using with the anisotropic ENMs by Riccardi *et al.*[67] and was used in other studies (e.g., Refs. 72–74). It is important to note that the 10 Å

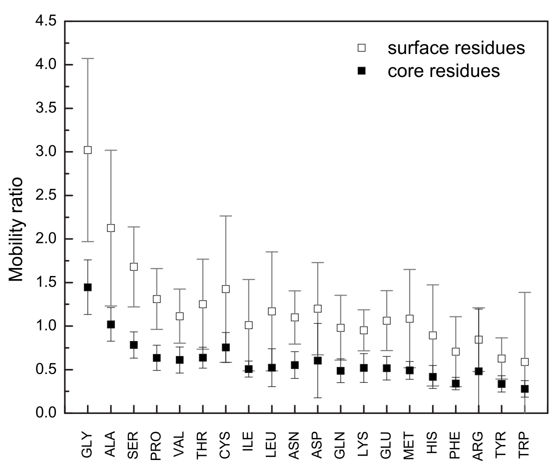

FIG. 3. The mobility ratios of surface and core residues. Surface (core) residues are defined as those residues with solvent accessible surface area higher(lower) than 25%. The error bars show the standard deviations.

cutoff results in a desired low level of the cutoff-related ruggedness of the protein-protein energy landscape,[75,76] which follows from the theory of minimally frustrated energy landscapes.[77]

B. The analysis of protein residue fluctuations

The analysis of the mobility ratios using Eq. (10) shows that large equilibrium fluctuations ($\mathcal{R} \geq 1$) of protein structures are associated with the oscillations of the center of mass of Gly, Ala, Ser, Pro, and Asp (group I) which are the most lightweight residues with the exception of Asp (Fig. 2). Modest fluctuations ($\mathcal{R}=0.7-1.0$; group II) are associated with six polar residues (Thr, Asn, Gln, Lys, Glu, Arg) and two nonpolar residues (Val, Cys). The small fluctuations ($\mathcal{R}=0.3-0.7$; group III) are associated with six nonpolar residues (Ile, Leu, Met, Phe, Trp) and polar residues His and Tyr. It is interesting to note that, with regards to hydrophilicity, groups I, II and III can be characterized as mixed, mostly polar and mostly nonpolar. Further we show that the groups occur due to the interplay between two factors determining the mobility ratios: the protein mass distribution and the core/surface locus of the residue.

Analysis of the scale of fluctuations of surface and core residues shows that on average all surface residues demonstrate larger fluctuations than the core residues (Fig. 3). Surface (core) residues are defined here as those residues which have relative solvent accessible surface area higher(lower) than 25% and are identified using NACCESS.[78] The difference is readily explained by the difference in numbers of nearest neighbors of surface and core residues (the environment effect). In comparison with the core residues, the surface residues have fewer nearest neighbors.[79] Therefore, they are less restricted and experience larger fluctuations. First reports of this effect go back to crystallographic studies of myoglobin[59,60] and lysozyme.[58] It has been shown that atomic mean-square displacements increase from the protein core to the protein surface. It was suggested[59] that, in general, proteins have a condensed core and a semiliquid surface. This was supported by the results of molecular dynamic simulations of carboxy myoglobin.[80] We would like also to note that the significant difference between mobility ratios of the most lightweight and most heavyweight residues (Fig. 2) as well as the nonmonotonic behavior for the midweight residues do not disappear if one considers core and surface residues separately (see Fig. 3). The likely reason for this is the interplay between the mass and packing effects. Indeed, in general, small residues are most-lightweight and maintain a smaller number of the atomic contacts with their structural neighbors than the larger residues, regardless of the residue position. Thus, they are less constrained and more mobile.

The same environment effect appears as a small root mean-square deviation between bound and unbound states of pocket side chains[81] or as a decreased number of rotamers allowable for buried amino-acids in comparison with the surface amino-acids.[82,83] This also clears up a seemingly striking difference in hydrophilicity between residues of Groups II and III. Indeed, amino acid residues are distributed nonhomogeneously in proteins. Polar residues prefer surface positions, but nonpolar residues are more often found in a protein core. That is why the mostly polar Group II demonstrates higher mobility ratios than the mostly nonpolar Group III. On the other side, high mobility ratios of nonpolar residues Gly and Ala suggest that the environment effect is not the only factor. The amplitude of fluctuations is inversely proportional to the effective amino acid masses [see Eq. (9)]. As a result, the largest fluctuations are associated with Gly and Ala, the most lightweight residues, but the smallest fluctuations are associated with Tyr and Trp, the most heavy residues (Fig. 2).

Average mobility ratios of the binding site (interface) and of other surface residues vary in intervals [0.4,2.6] and [0.6,3.0] correspondingly. Comparing fluctuations of the interface residues with other surface residues, we found that although, on average, interface is less mobile than the rest of the protein surface (Fig. 4), the noticeable difference ($\mathcal{R}_j^{\text{sur}} - \mathcal{R}_j^{\text{int}} > 0.25$, $\mathcal{R}_j^{\text{int,sur}}$ is the mobility ratio of the interface or other surface residue j) relates to Gly, Ala, Ser, Cys, Leu, and Trp. The average relative variance of the mobility ratios $1/6\Sigma_{j=1}^{6}(\mathcal{R}_j^{\text{sur}} - \mathcal{R}_j^{\text{int}})/\mathcal{R}_j^{\text{int}}$ of these residues is 39%. Standard errors of the average mobility ratios of the interface and noninterface surface residues vary in intervals [0.02,0.14] and [0.02,0.13] and are by the order of magnitude less than the average mobility ratios [see Supplementary Table II (Ref. 66)]. Four of these residues (Gly, Ala, Leu, and Ser) are the most common residues at protein interfaces, and residues Cys and Trp are the most infrequent interface residues.[7,8] The most conserved interface residue Trp (Ref. 9) also is the most stable one (see Fig. 4). Two other highly conserved interface residues (Met and Phe)[9] demonstrate decreased mobility in binding sites to a lesser extent. Note that the difference between the binding sites and the rest of the protein surface relates mainly to fluctuations of the nonpolar residues with the exception of Ser, a polar residue. These results are in agreement with the experimental observation of reduced fluctuations in binding sites of myoglobin[84] and bacteriorhodopsin[85] in comparison with fluctuations of the rest of macromolecules. Frauenfelder and McMahon[84] also noted that four (Leu29, Phe43, Val68, and Ile107) of the six residues with reduced fluctuations surrounding the oxygen molecule are nonpolar. The two other residues are His64 and His93 ($\mathcal{R}_{\text{His}}^{\text{sur}} - \mathcal{R}_{\text{His}}^{\text{int}} = 0.16$). The solvent-mediated attraction between nonpolar residues of a receptor and a ligand results in the hydrophobic contribution to binding free energy, which is considered to be one of the major factors stabilizing protein-protein complexes.[86,87] We suppose that Gly, Ala, Ser, Cys, Leu, and Trp form low-mobility surface "pads" that constitute a "landing ground" for binding proteins.

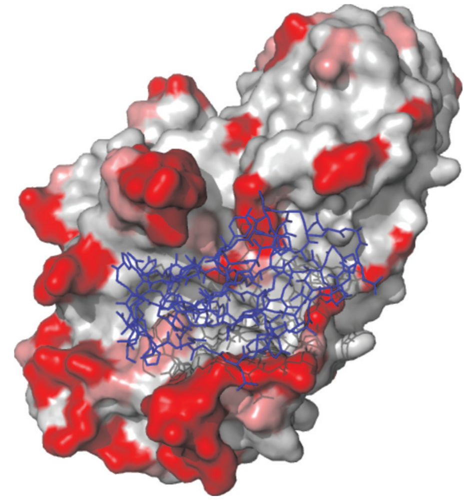

FIG. 5. The crystal structure of porcine pancreatic α-amylase in complex with the microbial inhibitor Tendamistat (blue) (Ref. 88). Highly fluctuating ($\mathcal{R} \geq 1$), moderately fluctuating ($R = 0.7-1.0$) and weakly fluctuating ($\mathcal{R} < 0.7$) residues of the hydrolase are in red, pink, and gray.

Figure 5 illustrates high stability (gray surface) of a binding groove in the porcine pancreatic α-amylase. The groove contains three catalytic residues Asp197, Glu233, and Asp300, which showed low mobility ratios. The binding site is surrounded by highly and moderately fluctuating convex areas, which is in agreement with the environment effect discussed above.

The larger ability to fluctuate of Group I residues provides an insight into the inability of sequences abundant in Gly, Ala, Ser, Pro, and Asp to fold into regular protein secondary structure elements (α-helices or β-strands). High mobility prevents the formation of long-range order, thus contributing to irregular protein secondary structure elements (loops). We computed the correlation coefficient between the mobility ratios and corresponding percentages of amino-acid residues in the data set of loops[2] (see Fig. 6). The analysis showed significant correlation with 0.9 correlation coefficient.

We suggest that the same reasoning explains features of amino-acid distributions observed in chameleon sequences[17,19,20] and disordered proteins.[22,23] Indeed, highly and moderately fluctuating amino-acid residues (in particular, Gly, Ala, Ser, Glu, and Lys) are abundant in disordered and "dual personality" protein fragments, whereas the residues with the low mobility ratio (e.g., Tyr, Trp, Phe, and Ile) are rarely found there.[22,23]

Statistics of protein residues in chameleon sequences show that Ala, Ile, Leu, and Val are the most frequent residues in chameleon sequences.[17,19] Since only Ala belongs to the Group I of highly fluctuating residues (Fig. 1), we can hypothesize that an instability driving helix$\leftrightarrow$sheet transi-

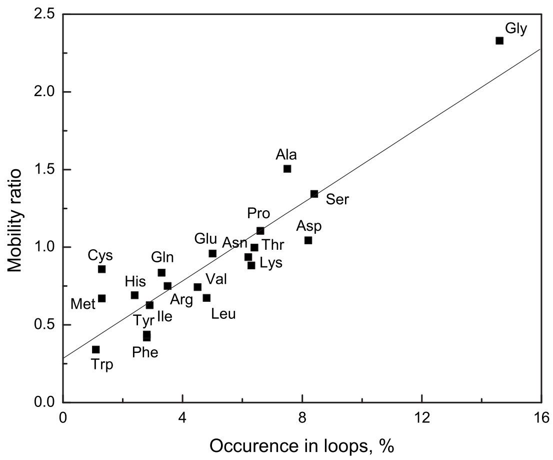

FIG. 6. The mobility ratios of protein residues vs their percentage compositions in protein loops. The correlation coefficient between the mobility ratios and the percentage compositions is 0.9 (Ref. 2).

tions may often originate at Ala residues if the other highly fluctuating residues are absent. Frequencies of occurrence of Gly and Ser residues increase with the increase of the length of the sequence.[17] Thus, in general, chameleon sequences may have several islands of instability. By exciting these islands locally (e.g., by mutations that change interactions of the islands with the rest of the protein or by ligands bound in the vicinity of the chameleon sequence), one could trigger a helix$\leftrightarrow$sheet transition. Mutations of a chameleon sequence, that change the mobility ratio of a sequence position significantly, can also provoke such transitions. It has been reported that a single mutation from Pro to Ala ($\mathcal{R}_{\text{Ala}}-\mathcal{R}_{\text{Pro}}=0.4$) converts a β-strands into an α-helix.[18] Mutations of two consecutive residues from Phe28Phe29 to Pro28Ile29 ($\mathcal{R}_{\text{Pro}}-\mathcal{R}_{\text{Phe}}=0.7$, $\mathcal{R}_{\text{Ile}}-\mathcal{R}_{\text{Phe}}=0.2$) converts an α-helix into a β-strand.[21]

The residue fluctuations derived by Eq. (9) increase with the increase of temperature. Therefore, we could expect that at the very early stages of protein thermal denaturation amino acid residues of the enhanced ability to fluctuate (Group I) and their structural neighbors will form first seeds of the unfolded phase. Since the majority of Group I amino acids (Gly, Ser, Pro, and Asp) shows higher propensities for loops than for helices or sheets,[1] it is possible that the nucleation of the unfolded phase starts in protein loops. Due to the increased ability to fluctuate, Group I residues can also be involved more often than other residues in equilibrium local folding-unfolding reactions scattered over the protein surface.[89,90]

The estimates of the residues' ability to fluctuate developed in this study are most likely to improve predictions of protein flexibility. Our results solve a longstanding contradiction of Vihinen *et al.*[91] classification that puts Gly, "generally considered to be the most flexible amino acid," in the middle of the flexibility scale. Although the Vihinen *et al.* classification is widely used,[22,23] this contradiction was noted by the authors[91] and others.[22] In the residue, classification based on our results Gly has the highest ability to fluctuate. The enhanced mobility of Gly has been commonly associated with the lack of a side chain, and thus greater conformational flexibility. The model developed here does not involve explicit side chains. However, it is able to reproduce the mobile character of Gly. This again points to the role of the two factors—the residue mass and the inter-residue contacts—in protein flexibility within the framework of the ENM and encourages using the model in the studies of the Gly-rich proteins and fragments (e.g., collagen, HIV-1 protease flaps,[92,93] or structural hinges[94]).

IV. CONCLUSIONS

The current work focuses on the fundamental relationship between the protein sequence, ability to fluctuate, and functionality of protein structures. We have considered the relationship within a framework of a novel elastic network model that allows accounting for the distribution of interatomic interactions within a coarse-grained approach. The model modifies a commonly used form of the Tirion potential with a spring constant proportional to the number of interatomic contacts between residues. We demonstrated that two factors, a protein mass distribution and a residue environment, determine the scale of fluctuations. The surface residues undergo larger fluctuations than the core residues in agreement with experimental observations.[58–60] On average, the protein interface is less mobile than the rest of the protein surface and contains low-mobility pads associated mainly with the nonpolar residues. We hypothesize that the conformational instability of protein loops, chameleon sequences, and disordered proteins relates to the high content of highly mobile residues and the lack of weakly fluctuating residues. The results show high correlation between fluctuations and the sequence composition of protein loops. Analysis of residue fluctuations and their propensities in secondary structure elements allows one to conclude that upon thermal denaturation the nucleation of the unfolded phase proceeds from protein loops. The results provide insight into structural fluctuations of proteins and facilitate better understanding of protein association mechanisms.

ACKNOWLEDGMENTS

The study was supported by Grant No. R01 GM074255 from NIH.

[1] S. Costantini, G. Colonna, and A. M. Facchiano, Biochem. Biophys. Res. Commun. **342**, 441 (2006).
[2] J.-M. Kwasigroch, J. Chomilier, and J. P. Mornon, J. Mol. Biol. **259**, 855 (1996).
[3] J. F. Leszczynski and G. D. Rose, Science **234**, 849 (1986).
[4] M. Levitt, Biochemistry **17**, 4277 (1978).
[5] K. T. O'Neil and W. F. DeGrado, Science **250**, 646 (1990).
[6] J. S. Richardson and D. C. Richardson, Science **240**, 1648 (1988).
[7] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, Proteins **43**, 89 (2001).
[8] O. Keskin, I. Bahar, A. Y. Badretdinov, O. B. Ptitsyn, and R. L. Jernigan, Protein Sci. **7**, 2578 (1998).
[9] B. Ma, T. Elkayam, H. Wolfson, and R. Nussinov, Proc. Natl. Acad. Sci. U.S.A. **100**, 5772 (2003).
[10] Y. Ofran and B. Rost, J. Mol. Biol. **325**, 377 (2003).
[11] B. Rost, J. Struct. Biol. **134**, 204 (2001).
[12] N.-V. Buchete, J. E. Straub, and D. Thirumalai, Curr. Opin. Struct. Biol. **14**, 225 (2004).
[13] C. Zhang, S. Liu, H. Zhou, and Y. Zhou, Protein Sci. **13**, 400 (2004).
[14] G.-Y. Chuang, D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda,

Biophys. J. **95**, 4217 (2008).
[15] P. Pfeffer and H. Gohlke, J. Chem. Inf. Model. **47**, 1868 (2007).
[16] A. M. Ruvinsky and A. V. Kozintsev, Proteins **58**, 845 (2005).
[17] M. Mezei, Protein Eng. **11**, 411 (1998).
[18] W.-Z. Yang, T.-P. Ko, H. S. Yuan, L. Corselli, and R. C. Johnson, Protein Sci. **7**, 1875 (1998).
[19] J.-T. Guo, J. W. Jaromczyk, and Y. Xu, Proteins **67**, 548 (2007).
[20] I. B. Kuznetsov and S. Rackovsky, Protein Sci. **12**, 2420 (2003).
[21] H. Tidow, T. Lauber, K. Vitzithum, C. P. Sommerhoff, R. Rösch, and U. C. Marx, Biochemistry **43**, 11238 (2004).
[22] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, J. Mol. Graphics Modell. **19**, 26 (2001).
[23] Y. Zhang, B. Stec, and A. Godzik, Structure (London) **15**, 1141 (2007).
[24] J. Liu, J. R. Faeder, and C. J. Camacho, Proc. Natl. Acad. Sci. U.S.A. **106**, 19819 (2009).
[25] G. A. Petsko and D. Ringe, Annu. Rev. Biophys. Bioeng. **13**, 331 (1984).
[26] O. Miyashita, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **100**, 12570 (2003).
[27] W. Zheng, B. R. Brooks, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **103**, 7664 (2006).
[28] W. Zheng and D. Thirumalai, Biophys. J. **96**, 2128 (2009).
[29] C. Xu, D. Tobi, and I. Bahar, J. Mol. Biol. **333**, 153 (2003).
[30] P. Maragakis and M. Karplus, J. Mol. Biol. **352**, 807 (2005).
[31] J.-W. Chu and G. A. Voth, Biophys. J. **93**, 3860 (2007).
[32] K. Hinsen, Proteins **33**, 417 (1998).
[33] M. K. Kim, G. S. Chirikjian, and R. L. Jernigan, J. Mol. Graphics Modell. **21**, 151 (2002).
[34] J. A. Kovacs, P. Chacón, and R. Abagyan, Proteins **56**, 661 (2004).
[35] F. Tama and C. L. Brooks III, J. Mol. Biol. **345**, 299 (2005).
[36] W. Zheng, B. R. Brooks, and G. Hummer, Proteins **69**, 43 (2007).
[37] D. Tobi and I. Bahar, Proc. Natl. Acad. Sci. U.S.A. **102**, 18908 (2005).
[38] D. Schneidman-Duhovny, R. Nussinov, and H. J. Wolfson, Proteins **69**, 764 (2007).
[39] S. E. Dobbins, V. I. Lesk, and M. J. E. Sternberg, Proc. Natl. Acad. Sci. U.S.A. **105**, 10390 (2008).
[40] K. Hinsen, Bioinformatics **24**, 521 (2007).
[41] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).
[42] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[43] H. Gohlke and M. F. Thorpe, Biophys. J. **91**, 2115 (2006).
[44] D. A. Kondrashov, Q. Cui, and G. N. Phillips, Jr., Biophys. J. **91**, 2760 (2006).
[45] K. Moritsugu and J. C. Smith, Biophys. J. **93**, 3460 (2007).
[46] A. May and M. Zacharias, Proteins **70**, 794 (2008).
[47] C. N. Cavasotto, J. A. Kovacs, and R. A. Abagyan, J. Am. Chem. Soc. **127**, 9632 (2005).
[48] B. K. Poon, X. Chen, M. Lu, N. K. Vyas, F. A. Quiocho, Q. Wang, and J. Ma, Proc. Natl. Acad. Sci. U.S.A. **104**, 7869 (2007).
[49] L.-W. Yang, E. Eyal, C. Chennubhotla, J. Jee, A. M. Gronenborn, and I. Bahar, Structure (London) **15**, 741 (2007).
[50] L. Yang, G. Song, A. Carriquiry, and R. L. Jernigan, Structure (London) **16**, 321 (2008).
[51] B. Erman, Biophys. J. **91**, 3589 (2006).
[52] K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **103**, 11844 (2006).
[53] F. Tama, F. X. Gadea, O. Marques, and Y.-H. Sanejouand, Proteins **41**, 1 (2000).
[54] D. T. Mirijanian and G. A. Voth, Proc. Natl. Acad. Sci. U.S.A. **105**, 1204 (2008).
[55] V. Tozzini, W. Rocchia, and A. J. McCammon, J. Chem. Theory Comput. **2**, 667 (2006).
[56] M. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).
[57] Y. Gao, D. Douguet, A. Tovchigrechko, and I. A. Vakser, Proteins **69**, 845 (2007).
[58] P. J. Artymiuk, C. C. F. Blake, D. E. P. Grace, S. J. Oatley, D. C. Phillips, and M. J. E. Sternberg, Nature (London) **280**, 563 (1979).
[59] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou, Nature (London) **280**, 558 (1979).
[60] H. Frauenfelder and G. A. Petsko, Biophys. J. **32**, 465 (1980).
[61] B. Halle, Proc. Natl. Acad. Sci. U.S.A. **99**, 1274 (2002).
[62] L. D. Landau and E. M. Lifshitz, *Mechanics* (Pergamon, New York, 1976), pp. 65–70.
[63] M. Levitt, C. Sander, and P. S. Stern, J. Mol. Biol. **181**, 423 (1985).
[64] D. A. Case, Curr. Opin. Struct. Biol. **4**, 285 (1994).
[65] N. Gō and H. A. Scheraga, J. Chem. Phys. **51**, 4751 (1969).
[66] See supplementary material at http://dx.doi.org/10.1063/1.3498743 for the correlation coefficients between B-factors of Ca atoms and calculated fluctuations of the residue centers of mass as a function of the cutoff (Table 1), and the standard deviations and standard errors of the average mobility ratios of interface and noninterface surface residues (Table 2).
[67] D. Riccardi, Q. Cui, and G. N. Phillips, Jr., Biophys. J. **96**, 464 (2009).
[68] B. J. Grant, A. A. Gorfe, and J. A. McCammon, PLOS Comput. Biol. **5**, e1000325 (2009).
[69] J. Qvist, M. Davidovic, D. Hamelberg, and B. Halle, Proc. Natl. Acad. Sci. U.S.A. **105**, 6296 (2008).
[70] S. Yang, N. K. Banavalib, and B. Roux, Proc. Natl. Acad. Sci. U.S.A. **106**, 3776 (2009).
[71] T. Z. Sen and R. L. Jernigan, in *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems, Optimizing the Parameters of the Gaussian Network Model for ATP-Binding Proteins*, edited by I. Bahar and Q. Cui (Chapman and Hall, London/CRC, Boca Raton, FL, 2006), pp. 171–186.
[72] B. Juanico, Y.-H. Sanejouand, F. Piazza, and P. D. L. Rios, Phys. Rev. Lett. **99**, 238104 (2007).
[73] F. Piazza and Y.-H. Sanejouand, Phys. Biol. **6**, 046014 (2009).
[74] W. Zheng, Proteins **76**, 747 (2009).
[75] A. M. Ruvinsky and I. A. Vakser, Bioinformatics **25**, 1132 (2009).
[76] A. M. Ruvinsky and I. A. Vakser, Proteins **70**, 1498 (2008).
[77] C.-J. Tsai, S. Kumar, B. Ma, and R. Nussinov, Protein Sci. **8**, 1181 (1999).
[78] S. J. Hubbard and J. M. Thornton, *NACCESS', Computer Program* (Department of Biochemistry and Molecular Biology, University College London, London, 1993).
[79] M. Gerstein and C. Chothia, Proc. Natl. Aca. Sci. U.S.A. **79**, 10167 (1996).
[80] D. Vitkup, D. Ringe, G. A. Petsko, and M. Karplus, Nat. Struct. Biol. **7**, 34 (2000).
[81] X. Li, O. Keskin, B. Ma, R. Nussinov, and J. Liang, J. Mol. Biol. **344**, 781 (2004).
[82] G. R. Smith, M. J. E. Sternberg, and P. A. Bates, J. Mol. Biol. **347**, 1077 (2005).
[83] D. Rajamani, S. Thiel, S. Vajda, and C. J. Camacho, Proc. Natl. Acad. Sci. U.S.A. **101**, 11287 (2004).
[84] H. Frauenfelder and B. McMahon, Proc. Natl. Acad. Sci. U.S.A. **95**, 4795 (1998).
[85] V. Rèat, H. Patzelt, M. Ferrand, C. Pfister, D. Oesterhelt, and G. Zaccai, Proc. Natl. Acad. Sci. U.S.A. **95**, 4970 (1998).
[86] D. Chandler, Nature (London) **437**, 640 (2005).
[87] I. A. Vakser and C. Aflalo, Proteins **20**, 320 (1994).
[88] G. Wiegand, O. Epp, and R. Huber, J. Mol. Biol. **247**, 99 (1995).
[89] Y. Bai, T. R. Sosnick, L. Mayne, and S. W. Englander, Science **269**, 192 (1995).
[90] B. Fierz, A. Reiner, and T. Kiefhaber, Proc. Natl. Acad. Sci. U.S.A. **106**, 1057 (2009).
[91] M. Vihinen, E. Torkkila, and P. Riikonen, Proteins **19**, 141 (1994).
[92] D. Hamelberg and J. A. McCammon, J. Am. Chem. Soc. **127**, 13778 (2005).
[93] V. Hornak, A. Okur, R. C. Rizzo, and C. Simmerling, Proc. Natl. Acad. Sci. U.S.A. **103**, 915 (2006).
[94] S.C. Flores, L.J. Lu, J. Yang, N. Carriero, and M.B. Gerstein, BMC Bioinf. **8**, 167 (2007).