Analysis of Correlations between Energy and Residue Fluctuations in Native Proteins and Determination of Specific Sites for Binding

Turkan Haliloglu[1] and Burak Erman[2]

[1]Polymer Research Center, Bogazici University, Bebek, Istanbul, Turkey
[2]Department of Chemical and Biological Engineering, Koc University, Sariyer, 34450 Istanbul, Turkey

(Received 21 October 2008; published 27 February 2009)

The Gaussian network model is used to derive the correlations between energy and residue fluctuations in native proteins. Residues are identified that respond strongly to energy fluctuations and that display correlations with the remaining residues of the protein at the highest modes. We postulate that these residues are located at specific sites for drug binding. We test the validity of this postulate on a data set of 33 structurally distinct proteins in the unbound state. Detailed results are presented for drug binding to the HIV protease.

DOI: 10.1103/PhysRevLett.102.088103 PACS numbers: 87.14.E−

Fluctuations of residues are now known to relate to the function of native proteins. Recognition, catalysis, binding, etc., are all affected in some way by these fluctuations. In the simplest picture, fluctuations are visualized as resulting from coupled harmonic motions of the residues from their mean positions [1–4]. Almost all of the work in this field has been confined to the mechanistic aspects of the phenomenon. In the present Letter, we emphasize the statistical thermodynamics basis of these motions with specific focus on correlations between energy and residue fluctuations. Energy fluctuations in a protein may result from the instantaneous fluctuations of residue positions or from external sources that are transferred through the surface of the protein. We ask and try to answer the question of which residues are affected most when a fluctuation in the energy of the system takes place. Using a statistical mechanical model, we extend the Gaussian network model (GNM) [1] and show that the residues that are strongly affected by energy fluctuations, which we call "highly excitable residues," are correlated with a large number of the remaining residues of the protein at the highest modes. In other words, these residues are part of the network of residues whose fluctuations are strongly coupled with each other. We are particularly interested in the highest modes since they reflect local events at the residue level, while the lower modes reflect the global motions of the protein [5,6]. Identifying a residue whose fluctuations are coupled with most of the residues in the protein is important for understanding processes that involve energy exchange. We postulate that these residues are located at specific sites for binding of natural substrates or drugs. We test this postulate on a data set of 33 structurally distinct proteins in the unbound state and show that the most excitable residues at the highest modes are located in the vicinity of the sites where substrate binding takes place. In our data set, the substrates are drug molecules.

Correlations between fluctuations of residues can easily be determined by the GNM. In order to introduce correlations between energy and residue displacements, we rederive the GNM using a statistical mechanics model that is based on evaluating the probability distribution of the instantaneous energy and positions of residues. We consider the protein to be in diathermal, pressure (P) and a force ($\boldsymbol{F}$) reservoir, as a result of which the energy, the volume, and the positions of residues exhibit fluctuations. The distribution function f is given by

$$f(\hat{U},\hat{V},\hat{\boldsymbol{R}}) = \exp\left\{-k^{-1}\left(S-\frac{U}{T}-\frac{P}{T}V+\frac{\boldsymbol{F}}{T}\cdot\boldsymbol{R}\right) - k^{-1}\left(\frac{\hat{U}}{T}+\frac{P}{T}\hat{V}-\frac{\boldsymbol{F}}{T}\cdot\hat{\boldsymbol{R}}\right)\right\}, \tag{1}$$

where S, U, and V are the thermodynamic (mean) entropy, energy, and volume, respectively, and $\hat{S}$, $\hat{U}$, and $\hat{V}$ are their instantaneous values. T is the temperature. k is the Boltzmann constant. The correlation of fluctuations of the ith and jth residues is obtained from

$$\langle\Delta\mathbf{R}_i\Delta\mathbf{R}_j^T\rangle = \sum(\hat{\mathbf{R}}_i-\mathbf{R}_i)(\hat{\mathbf{R}}_j-\mathbf{R}_j)^T f(\hat{U},\hat{V},\hat{\mathbf{R}}). \tag{2}$$

Here the superscript T denotes transpose, the summation is over all allowable states, and $\Delta\boldsymbol{R}$ is the column vector of fluctuations of residues. Using Eq. (1) in Eq. (2) and following the derivation given on pp. 426–427, Eqs. (9)–(14) and (19), of Ref. [7], leads to

$$\langle\Delta\mathbf{R}_i\Delta\mathbf{R}_j^T\rangle = kT\left(\frac{\partial\mathbf{R}_i}{\partial\mathbf{F}_j}\right). \tag{3}$$

This equation relates the correlation of fluctuations of residues to the mean position vectors and the force. For the special case of the GNM, the relation between the force and the residue positions is linear: $\boldsymbol{F}=\boldsymbol{\Gamma}\boldsymbol{R}$, where $\boldsymbol{\Gamma}$ is the matrix of force constants, defined as

$$\Gamma_{ij} = \begin{cases} -\gamma^* & i\neq j \quad \text{and} \quad R_{ij}\leq r_{\text{cutoff}}, \\ 0 & i\neq j \quad \text{and} \quad R_{ij}> r_{\text{cutoff}}, \\ -\sum_k \gamma^* & i=j\neq k. \end{cases} \tag{4}$$

Here R_{ij} is the distance between residue i and j, and r_{cutoff} is the distance that defines the neighborhood condition generally taken between 6.5–7 Å. $\gamma*$ is a scaling parameter.

In the GNM approximation, Eq. (3) reduces to

$$\langle \Delta \boldsymbol{R}_i \Delta \boldsymbol{R}_j^T \rangle = kT\boldsymbol{\Gamma}^{-1}. \tag{5}$$

The correlation $\langle \Delta U \Delta \boldsymbol{R}_i \Delta \boldsymbol{R}_j^T \rangle$ between energy fluctuations and those of the fluctuations of the positions of the ith and jth residues is defined by

$$\langle \Delta U \Delta \mathbf{R}_i \Delta \mathbf{R}_i^T \rangle = \sum (\hat{U} - U)(\hat{\mathbf{R}}_i - \mathbf{R}_i) \times (\hat{\mathbf{R}}_j - \mathbf{R}_j)^T f(\hat{U}, \hat{V}, \hat{\mathbf{R}}). \tag{6}$$

Using Eq. (1) in Eq. (6) leads to (see Ref. [7])

$$\langle \Delta U \Delta \mathbf{R}_i \Delta \mathbf{R}_j^T \rangle = (kT)^2 \left\langle \frac{\partial^2 U}{\partial \mathbf{F}_j \partial \mathbf{F}_i} \right\rangle. \tag{7}$$

Performing the differentiation shown in Eq. (7) and using the relations $\frac{\partial}{\partial \boldsymbol{F}_j}\left(\frac{\partial U}{\partial \boldsymbol{F}_i}\right) = \frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{F}_j} = \boldsymbol{\Gamma}^{-1}$ leads to the expression

$$\langle \Delta U \Delta \mathbf{R}_i \Delta \mathbf{R}_j^T \rangle = (kT)^2 (\boldsymbol{\Gamma}^{-1})_{ij} = kT \langle \Delta \mathbf{R}_i \Delta \mathbf{R}_j^T \rangle. \tag{8}$$

Thus, fluctuations of energy are distributed to the residues in proportion to the correlations of fluctuations. The diagonal elements $\langle \Delta \mathbf{R}_i \Delta \mathbf{R}_j^T \rangle$ are positive by definition. Therefore, the terms $\langle \Delta U (\Delta \mathbf{R}_i)^2 \rangle$ are positive. This means, for a given residue i, for example, that a positive value of ΔU couples with large values of $(\Delta R_i)^2$ and a negative ΔU couples with small values of $(\Delta R_i)^2$. This follows from the definition of the correlation obtained by averaging over k time points: $\langle \Delta U (\Delta \mathbf{R}_i)^2 \rangle = \frac{1}{k} \sum_{j=1}^{k} \Delta U(t_j)[\Delta \mathbf{R}_i(t_j)]^2$. For the off-diagonal terms, the same pattern holds. If $\langle \Delta \mathbf{R}_i \Delta \mathbf{R}_j^T \rangle > 0$, then positive energy fluctuations pick up the large positive $\Delta \mathbf{R}_i \Delta \mathbf{R}_j^T$'s. Conversely, if $\langle \Delta \mathbf{R}_i \Delta \mathbf{R}_j^T \rangle < 0$, then positive energy fluctuations pick up the large negative $\Delta \mathbf{R}_i \Delta \mathbf{R}_j^T$'s. Equation (8) may be regarded in two equivalent ways: The fluctuations in the energy of the protein drives the fluctuations, or, reciprocally, the fluctuations in residue positions result in fluctuations of the protein energy. Derivation of Eq. (8) is given in the supplementary material [8].

Equation (8) shows how residues are excited by energy fluctuations. A physically more relevant situation is to find how energy fluctuations affect the distance between two residues. We define the mean-square fluctuations $\langle \Delta R_{ij}^2 \rangle \equiv \langle (\Delta \boldsymbol{R}_i - \Delta \boldsymbol{R}_j)^2 \rangle$ of the distance between residues i and j. With this, Eq. (8) takes the form

$$\begin{aligned} \langle \Delta U (\Delta \mathbf{R}_{ij})^2 \rangle &= (kT)^2 [(\boldsymbol{\Gamma}^{-1})_{ii} - 2(\boldsymbol{\Gamma}^{-1})_{ij} + (\boldsymbol{\Gamma}^{-1})_{jj}] \\ &= kT \langle (\Delta \mathbf{R}_{ij})^2 \rangle. \end{aligned} \tag{9}$$

Equation (9) shows that the coupling between energy and distance fluctuations is directly proportional to the corresponding distance fluctuations.

Our calculations show that if a residue i can be excited at the highest modes and if it is coupled to the remaining residues j through Eq. (9), then it is highly probable that residue i is associated with a binding site. We verify our hypothesis by determining the drug binding sites of several protein-drug complexes and comparing with experimental results.

Results.—We analyzed a data set of 33 structurally distinct proteins in the unbound state [7]. These are proteins whose drug bound structures are also available, and thus drug binding sites are known *a priori*. For each unbound structure, the GNM calculations are performed to identify the residue i that displays the highest distance fluctuations $\langle \Delta R_{ij}^2 \rangle$ in the fastest modes of motion with the remaining residues j of the protein. To be specific, we look at the highest mode only. Results show that the residues identified in this manner are located in the known drug binding sites; these residues are highly excitable residues on the proteins for the drugs. In this Letter, we mainly elaborate on the HIV-1 protease structure. Results for the other 32 systems of the data set are presented in Table I. Additional information can be seen in the supplementary material [8].

HIV-1 protease is an aspartic protease, which is essential for the life cycle of HIV [9] and thus one of the main drug targets. The HIV protease exists as a homo dimer, shown in Fig. 1, with each subunit made up of 99 amino acids. The two segments Leu24–Glu34 and Asn83–Ile93 of each of the two monomers are highly protected. The active site has the Asp25-Thr26-Gly27 sequence where the two Asp25 residues, one from each chain, act as the catalytic residues [11]. The protease recognizes ten nonhomologous octameric nonsymmetric substrate sites within the virus's Gag and GagPol polyproteins at these active sites [12,13]. Thus, these regions are competitive sites for both the natural substrates and the drugs. In addition to its family-classifying conserved sequence of the active site, a second highly conserved sequence of Gly86-Arg87 is observed in the viral enzyme. While the active site triad Asp25-Thr26-Gly27 is common to all aspartic acid proteases, residues Gly86-Arg87-Asn/Asp88 are unique to retroviral proteases [14]. Arg87 is also a significant residue in dimerization [15,16].

We utilized an 11 ns molecular dynamics (MD) simulation of the HIV-1 protease, the coordinates of which are taken from the crystal structure of the HIV-substrate complex [Protein Data Bank (PDB) code: 1F7A] with the substrate removed. In this way, the molecule is relaxed from the crystal structure. We then performed GNM calculations on several snapshots taken from the MD trajectory, which may possibly reflect some conformational changes.

Figure 2(a) displays the mean-square distance fluctuation $\langle \Delta R_{ij}^2 \rangle$ in the fastest mode of motion calculated for one of the subunits (A) of the protease from the dimer structure. As seen, the residue pairs that display the highest distance

TABLE I. Key residues in drug binding predicted by the GNM. A more detailed list of residues is given in the supplementary material [8].

PDB code	Predicted Eq. (9)	Known
1A6U	92	93
1QIF	198–207	199, 200
3APP	30, 33, 121	31, 33, 121
1DJB	131–133	130, 132
1BYA	56–60, 63	53, 55
1CGE	194, 219, 222	196, 219, 222
1IFB	40, 49	40, 49
1A4J	35, 50, 97, 108	35, 50, 95, 100
1IME	197, 200, 221	195–197, 220
1NNA	116	118, 119
1AHC	154	155
2TGA	194, 196, 212	195, 213
1PHC	254, 257, 258	252, 253
1PSN	213–220, 303	215–219, 303
3LCK	319, 324, 371	319, 323, 371
1BRQ	138	133, 135
1BBS	213, 220, 300	213, 220, 300
1STN	41	41
1PTS	104, 106, 130	108, 128
2RTA	108	108
2CTB	145, 251, 253	145, 250, 253
2CBA	66, 93	64, 92, 94, 96
1KRN	64, 69	62–64, 71
2SIL	55–57, 65	56, 62
1L3F	163, 233–235	166, 231
1YPI	162–164, 230	165, 230, 232
1CHG	44, 198, 213	42, 195, 213
6INS	6	6
2PTN	194, 196, 197	195
3P2P	42, 44	45
5CPA	66, 70, 71	69, 71, 72
7RAT	108, 109	119, 120, 121

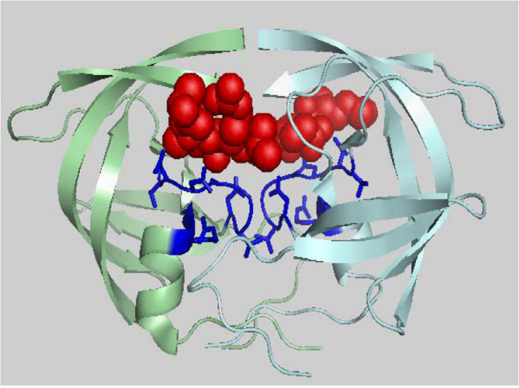

FIG. 1 (color online). Ribbon diagram of a HIV-1 protease-substrate complex structure (1F7A). Red spheres show the substrate residues, and blue sticks show the residues that touch the substrate and display high distance fluctuations with the remaining residues of the protein in the unbound state. This picture is prepared using Pymol [10].

fluctuations are (i) between Asn25-Thr31 that has the active site residues and (ii) Arg87 and its nearby residues. Here Arg87 is a unique residue that displays high–mean-square distance fluctuation correlations with all of the residues in the structure. The ten pairs of residues, out of possible 4802 pairs, with the highest distance fluctuations are ordered as: (Gly86, Arg87), (Arg87, Thr91), (Arg87, Leu90), (Thr31, Arg87), (Asn25, Arg87), (Thr26, Arg87), (Asp29, Arg87), (Asp30, Arg87), (Gly27, Arg87), and (Ala28, Arg87). Arg87 appears in all pairs as visualized also in Fig. 2(a). Among these, (Arg87, Thr91) and (Arg87, Leu90) are associated with dimerization, and all of the other pairs are involved in substrate or drug binding. These residues, which are also drug binding sites at the same time, are colored in blue in a protease-substrate complex structure (1F7A) in Fig. 1. Figure 2(b) displays the number of occurrences of each residue that is among the highly correlated pairs in the structure, that is, the number of occurrences of each residue in the top 2% of the total number of pairs that display the highest distance correlation values.

Substitution mutations of the highly conserved Arg87 are known to result in loss of proteolytic activity [17]. It is suggested that the highly conserved Arg87 of the HIV dimer is involved in ion pairing with the similarly highly conserved Asp29 to form the specific structure for substrate binding [18]. In a recent study [19] the analysis of five FDA-approved drugs suggested that improving the interactions between these drugs and residues Leu23, Ala28, Asp29, Gly49, and Arg87 in addition to Asp25 and Thr26, would possibly enhance the ability of the protease to combat drug resistance. Arg87, as also noted above, is a significant residue in dimerization.

When we consider the next fastest two modes, we observe that Ile84, Gly86, and Asn88-Thr91, which are the nearby residues of Arg87, display enhanced high distance fluctuations with several other residues of the structure. So

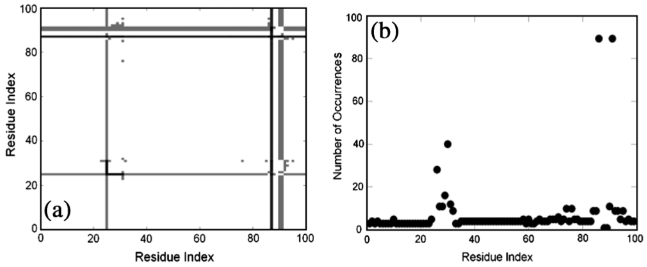

FIG. 2. (a) Contour map of $\langle \Delta \mathbf{R}_{ij}^2 \rangle$ versus i and j in the fastest mode for the monomer A of HIV-1 protease. Each positive value of $\langle \Delta \mathbf{R}_{ij}^2 \rangle$ corresponding to the ith residue along the abscissa and to the jth residue along the ordinate is denoted by a point on the map. Nearby points appear as horizontal and vertical continuous strips. Darker points correspond to larger values. (b) Number of occurrences of each residue in the top 2% of the total number of pairs that display the highest distance fluctuation correlation values $\langle \Delta \mathbf{R}_{ij}^2 \rangle$.

do Asn25 and Thr31. These are the few residues of the structure that could be excitable in the fastest modes of motion. Further, calculations performed on the dimer also display the same sites found for the monomer as the residues that couple strongly to the fluctuations in the energy.

The two sites that appear in the monomer and dimer calculations are the observed binding sites in several available substrate and drug bound complex structures. Additionally, the calculations performed on the unbound crystal structure (1HSI) also give the same result. This suggests the invariance of this behavior in different conformations. The conservation of these sites in the monomeric state is particularly intriguing, suggesting that, although this protease functions in the dimeric state, the fluctuation behavior of the functional sites of Leu24–Glu34 and Asn83–Ile93 are imprinted in the monomeric system; these are apparently the key functional sites of the structure.

The remaining 32 structures are analyzed similarly to the HIV-1 protease example given above. The results are presented in Table I. The first column identifies the PDB codes of the unbound (bound) proteins. The second column gives the residues i that have the highest $\langle \Delta U(\Delta \mathbf{R}_{ij})^2 \rangle$ values from Eq. (9) with the remaining residues j of the protein in the highest mode. The third column lists the known binding sites for drugs, taken from PDBsum [20], and is to be compared with the predictions given in the second column. There are also a few residues that are predicted by Eq. (9) but are not known as drug binding sites. For space reasons we present these in the extended form of Table I given in the supplementary material [8]. The fourth column of the extended table lists these additional residues obtained from Eq. (9) but not known as drug binding. However, several of these residues are either in contact with or in the close vicinity of the drug binding residues. Two structures, 4CA2 and 1PDY, in the original list [7] are excluded from analysis because the ligand in each of these two structures is not on the surface but located in a channel.

Discussion.—Comparison of columns 2 and 3 of Table I shows that residues that are highly correlated (high distance fluctuation correlation) with the remaining residues of the protein are either the drug binding sites or an immediate sequence or spatial neighbors of these sites. The positions of these residues are mostly at the cavities on the surface and sometimes more into the core of the structure.

To this end, it should also be noted that the strongly coupled fluctuations identify a network of interactions in the structure that could possibly be associated with the function. Recently, it was shown that the catalytic activity could be controlled by distal mutations to the catalytic sites in dihydrofolate reductase (DHFR) [21], and a network of residues was suggested for functional interactions of DHFR. Here it is possible to show that the residues in the proposed interaction network for DHFR overlap the residues that are identified also by the present method for DHFR (see supplementary material [8]).

The results of [22,23] implied the importance of the fluctuations in the fastest modes of motion in ligand-receptor and protein-protein interactions, in general. Here we present evidence that residues that exhibit large distance fluctuation correlations with the remaining residues of the protein in the fastest mode are associated with drug or substrate binding sites. Further, the identified binding sites are those that exhibit strong coupling with energy fluctuations. Shown with the drug-protein complex structures here, this could be a general phenomenon that has significance in binding mechanism.

We thank Ms. Ozen for the MD simulations of the HIV protease. T. H. acknowledges EU FP6-2004-ACC-SSA-2: Project No. 517991 and Boğaziçi University Funds: Project No. 08A505D.

[1] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[2] T. Haliloglu, I. Bahar, and B. Erman, Phys. Rev. Lett. **79**, 3090 (1997).
[3] M. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).
[4] Q. Cui and I. Bahar, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems* (Chapman and Hall, London, 2006).
[5] I. Bahar *et al.*, Phys. Rev. Lett. **80**, 2733 (1998).
[6] M. C. Demirel *et al.*, Protein Sci. **7**, 2522 (1998).
[7] A. T. R. Laurie and R. M. Jackson, Bioinformatics **21**, 1908 (2005).
[8] See EPAPS Document No. E-PRLTAO-102-069910 for (1) fluctuations of residues in the fastest mode versus drug binding sites, (2) the statistical thermodynamics equations of fluctuations, and (3) fluctuations of DHFR. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[9] I. T. Weber and R. W. Harrison, Protein Eng. **12**, 469 (1999).
[10] W. L. DeLano, The PyMOL Molecular Graphics System (2002), http://www.pymol.org
[11] A. Wlodawer and J. Erickson, Annu. Rev. Biochem. **62**, 543 (1993).
[12] K. C. Chou, Anal. Biochem. **233**, 1 (1996).
[13] L. E. Henderson *et al.*, J. Virol. **62**, 2587 (1998).
[14] L. H. Pearl and W. R. Taylor, Nature (London) **329**, 351 (1987).
[15] I. T. Weber, J. Biol. Chem. **265**, 10 492 (1990).
[16] J. M. Louis *et al.*, J. Biol. Chem. **278**, 6085 (2003).
[17] J. M. Louis *et al.*, Biochem. Biophys. Res. Commun. **164**, 30 (1989).
[18] T. Blundell and L. H. Pearl, Nature (London) **337**, 596 (1989).
[19] W. Wei Wang and P. A. Kollman, Proc. Natl. Acad. Sci. U.S.A. **98**, 14 937 (2001).
[20] http://www.ebi.ac.uk/pdbsum/.
[21] F. Wong, T. Selzer, and S. J. Benkovic *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 6807 (2005).
[22] T. Haliloglu, E. Seyrek, and B. Erman, Phys. Rev. Lett. **100**, 228102 (2008).
[23] A. Ertekin, R. Nussinov, and T. Haliloglu, Protein Sci. **15**, 2265 (2006).