Relationships between ligand binding sites, protein architecture and correlated paths of energy and conformational fluctuations

Burak Erman

Chemical and Biological Engineering Department, Koc University Istanbul, Turkey
E-mail: berman@ku.edu.tr

Received 26 April 2011
Accepted for publication 21 July 2011
Published 10 August 2011
Online at stacks.iop.org/PhysBio/8/056003

Abstract
The statistical thermodynamics basis of energy and residue position fluctuations is explained for native proteins. The protein and its surroundings are treated as a canonical system with emphasis on the effects of energy exchange between the two. Fluctuations of the energy are related to fluctuations of residue positions, which in turn are related to the connectivity matrix of the protein, thus establishing a connection between energy fluctuation pathways and protein architecture. The model gives the locations of hotspots for ligand binding and identifies the pathways of energy conduction within the protein. Results are discussed in terms of two sets of models, the BPTI and 12 proteins that contain the PDZ domain. A possible use of the model for determining functionally similar domains in a diverse set of proteins is pointed out.

1. Introduction
The native protein in the cell is not an isolated system but undergoes continual exchange of energy with its surroundings, as a result of which the protein performs its function. The fluctuations of energy act as the driving potential behind several phenomena affecting the function of the protein, notably the spatial fluctuations of the residue positions. The latter are significant in native proteins at physiological temperatures, as evidenced by experimental B-factors. The amplitudes of fluctuations are spatially inhomogeneous, being different for different residues and exhibiting an inverse dependence on the number of spatial neighbors of each residue, thus establishing an important connection to the topology of the native state, or the contact map of the protein. The present study aims to explain the statistical thermodynamics basis of the relationship between energy and conformational fluctuations of the protein at the residue level. A deeper understanding of the thermostatistics of the protein allows us to answer, or at most consider, several questions that have recently been attracting the attention of investigators on the structural basis of energetic interactions upon ligand binding (Lockless and Ranganathan 1999, Freire 1999, Suel et al 2003). Specifically, we address the question of the partitioning of the instantaneous increment of energy from the surroundings among the various residues, and the relation of this partitioning to the topology of the three-dimensional structure. The analysis shows, as will be discussed in detail in section 3, that the exchange of energy between the protein and the surroundings is not spatially isotropic nor random. The surroundings may be water molecules as well as ligands, cell wall, DNA, proteins, etc. Certain residues, which we recently called the ‘energy gates’ (Tuzmen and Erman 2011), are the hotspots that play major role in energy exchange with the surroundings. These are residues that can respond to the incoming energy and can share it with others in the protein, compared to several other residues at the surface that show negligible response to energy perturbations. This observation has immediate and important consequences regarding the ligand binding problem. The analysis also shows that the energy taken up from the surroundings does not diffuse randomly within the three-dimensional structure but rather...
follows specific paths which we recently called ‘interaction pathways’ (Haliloglu et al 2010, Tuzmen and Erman 2011). The relation of this observation to allosteric interactions of proteins, on which a large number of studies now converge, is obvious (Monod and Wyman 1965).

Interest in energy fluctuations in proteins is not new, of course, since it is directly related to the heat capacity, $C_P$, by the relation $C_P = \langle \Delta U^2 \rangle / kT^2$, which has been given most transparently in the work of Prabhu and Sharp (Prabhu and Sharp 2005) on which we elaborate further in section 2.1 below. Interest on energy fluctuations has been mostly limited to the study of changes in the heat capacity relating to events during folding or unfolding. Following the pioneering works of Cooper (1976) and Sturtevant (1977), several papers have been written on this subject and reviewed extensively in Prabhu and Sharp (2005). The sources of heat capacity changes with increasing temperature such as formation of cages of structured water around nonpolar groups, the breaking of hydrogen bonds and the increase of internal vibrational degrees of freedom have been studied in detail both theoretically and experimentally (Prabhu and Sharp 2005, Sturtevant 1977). In the present paper, we do not address the issue of heat capacity changes but instead focus on the partitioning of the fluctuations at the residue level, i.e. the relationship between energy fluctuations and the vibrational degrees of freedom of individual residues.

Several papers form the background and the underlying material for the present study which are briefly reviewed here. The terminology and the formulation of fluctuations in chapter 19 of the classical book on thermostatistics by Callen (1985) are adopted. This terminology was used for proteins previously (Oylumluoglu et al. 2006, 2007) with emphasis on the electric field as the thermodynamic extensive variable and the protein dipole moment as the thermodynamic force. Electric field fluctuations are not considered here but can however be handled with the general expression of the probability function given in appendix A. The present work is motivated by the work of Piazza, de Los Rios and Sanejouand (2005) who pointed out that the energy of a protein is dissipated to the environment only by surface atoms, and bulk atoms exchange energy with each other only. A simplified solution of the Fokker–Planck equation based on harmonic potentials, energy given to a mode of the system remains there indefinitely, and it is only the anharmonicities in the potential that allow the energy to flow to different modes. It was later shown by the pioneering molecular dynamics simulations of Moritsugu, Miyashita and Kidera (2000) that through a third-order nonlinearity, the vibrational energy was transferred from a normal mode to a very few number of normal modes in a protein. When regarded as a canonical system, the modes of a protein, even in the absence of anharmonicities, may be excited by the surrounding liquid and a wide spectrum of energy relaxation may occur as pointed out before (Piazza et al 2005). In this respect, although anharmonicities are important, such as for forming discrete breathers (Juanico et al 2007, Piazza and Sanejouand 2008) for example, they are not essential for the discussion of energy fluctuations in proteins. Finally, the large repertoire of papers on the relation of energy fluctuations on allosteric reactions of proteins, reviewed by Swain and Gierasch (2006), is the relevant background to this work. Notably, the work of Ranganathan and collaborators (Lockless and Ranganathan 1999, Suel et al 2003) has broadened our perspective on allosteric communication and protein architecture. The results and predictions of the fluctuation model of proteins presented here have close bearing to their work. Along similar lines, the model developed by Freire and collaborators (1999) that emphasizes the propagation of binding interactions to remote sites in proteins depends heavily on the concept of energy fluctuations and their relation to protein structure.

The second aim of this paper is to recapitulate the statistical thermodynamics basis of the Gaussian network model (GNM) (Bahar et al 1997). This model has been derived with reference to the statistical mechanics of random Gaussian networks (Kloczkowski et al 1989) and has a partition function equivalent to that of a canonical system, but the significance of energy exchanges of the protein with its surroundings has not been visualized then. The latter is emphasized in the present study. In doing this, some of the familiar equations of the GNM are repeated, for the sole interest of explaining their connection to the canonical statistical mechanical system. More recent work on the thermostatistics of native proteins from our group elaborated on different
aspects of the problem such as anharmonic probability distribution of residue fluctuations (Kabackioglu et al 2010, Yogurtcu et al 2009), quasi-harmonic mode coupling (Gur and Erman 2010), binding interactions (Haliloglu and Erman 2009, Haliloglu et al 2008) and predicting interaction pathways (Haliloglu et al 2010, Tuzmen and Erman 2011). These topics are unified under a thermostatistics formalism in the present paper, and it is hoped that this theoretical framework will allow for future improvements of the model.

The paper is organized as follows. In section 2, the thermodynamic variables are defined in the entropy representation for the model. The protein plus its surroundings form an isolated system, i.e. a canonical system. Then, the harmonic approximation is adopted and the relationship between the energy and residue fluctuations is given. The main finding of the paper, correlations of energy fluctuations, is then presented in terms of two related matrices: the connectivity matrix of the protein as a graph and the gamma matrix of the GNM. The expression derived for energy fluctuation correlations establishes the relationship between energy fluctuations and protein architecture. Finally, the ‘heat capacity of the distance between two residues’ is introduced as a new concept which is a measure of the response of a pair of residues to energy fluctuations. All relevant thermostatistics information is summarized in six appendices. In the interest of giving a broad perspective to the thermostatistics of native proteins, the most general form of the probability distribution function for fluctuations is given in appendix A, which is then simplified for use in this study. Some of the appendices contain information which is already well known but introduced here to eliminate cross-referencing, and some of the appendices contain details of mathematical derivations of the equations given in the text. The predictions of the model are compared, for proof-of-concept purposes, for two widely studied systems: (i) bovine pancreatic trypsin inhibitor, BPTI, and (ii) 12 proteins of the PDZ domain.

2. Material and methods

2.1. The model

The model consists of a protein of N residues embedded in surroundings of \(N_s\) molecules. The protein and the surroundings form an isolated system. The pressure, temperature and total number of molecules are fixed. We assume that the protein is in a state of local energy minimum where small fluctuations away from mean positions are always restored back to the mean positions. The protein exchanges energy with the surroundings, resulting in fluctuations of the energies of the individual residues and of residue positions. In a general model, the volume of the protein also fluctuates, but we assume that changes in conformation leading to anisotropic fluctuations of shape are much larger than the changes in volume. We therefore assume the protein to be incompressible. We adopt the residue-based coarse graining approximation, where atoms of each residue are centered at the corresponding alpha carbon, \(C^α\). As was pointed out recently, the residue-level approximation yields the minimal set of thermodynamic variables needed to explain energy fluctuations in native proteins (Rader 2010).

2.2. Thermodynamic variables

Each residue is identified by its position vector, \(R_i\), where the subscript \(i\) identifies the residue. Without the subscript, \(R\) represents the set of all position vectors of the \(N\) residues. The thermodynamic extensive variables, the energy, \(U(R)\), and entropy, \(S(R)\), are functions of residue positions, with instantaneous values \(\hat{U}(R)\) and \(\hat{S}(R)\). We assign an entropy \(S_i(R)\) and an energy \(U_i(R)\) to each residue. The extensive nature of the entropy and energy requires that

\[
U = \sum_{i=1}^{N} U_i \quad S = \sum_{i=1}^{N} S_i. \tag{1}
\]

2.3. The entropy representation of the protein

The entropy of the \(i\)th residue may be written as a function of the energies of the constituent subsystems as

\[
S_i = S_i(U_j(R)) \quad j = 1, \ldots, N. \tag{2}
\]

Equation (2) may be inverted to yield the energy of each residue

\[
U_i = U_i(S_j(R)) \quad j = 1, \ldots, N. \tag{3}
\]

Using the first of equation (1), the total energy of the protein is written in the entropy representation as

\[
U(S) = \sum_{i=1}^{N} U_i(S_j(R)) \quad j = 1, \ldots, N. \tag{4}
\]

2.4. The system as a canonical ensemble

The protein exchanges energy with its surroundings and the system constitutes a canonical ensemble. The probability \(f(\hat{U}(R))\) that the protein has the instantaneous energy \(\hat{U}(R)\) follows from the general expression (A.1) as

\[
f(\hat{U}(R)) = \exp \left\{ \frac{1}{kT} (A - \hat{U}) \right\}, \tag{5}
\]

where \(A = U - TS\) is the Helmholtz free energy. The probability of a fluctuation \(\Delta \hat{U} = \hat{U} - U\) of the energy is obtained from equation (5) as

\[
f(\Delta \hat{U}) = \exp \left\{ -\frac{S}{k} \right\} \exp \left( -\frac{1}{kT} \Delta \hat{U} \right). \tag{6}
\]

2.5. Relationship between energy and residue position fluctuations and the force–fluctuation relation

The contributions to the fluctuating energy of a residue coming from the surroundings of that residue may be due to hydrogen bonding, Lenard-Jones-type forces, dipolar coupling, electrostatic coupling, covalent bonding, etc. Whatever the source of the energy is, its fluctuations are coupled with the spatial fluctuations \(\Delta \hat{R}\) of the residues. In the simplest approximation, we assume that residues \(i\) and \(j\) that are within the interaction distance of each other interact
with a harmonic potential. The fluctuation of the energy $\Delta U_{ij}$ in this case is related to the residue position fluctuations $\Delta \hat{R}_i$ and $\Delta \hat{R}_j$ by (Hinsen 1998)

$$\Delta U_{ij} = [k_{ij} \cos^2(\alpha_{ij})](\Delta \hat{R}_j - \Delta \hat{R}_i)^2,$$

(7)

where $k_{ij}$ is the spring constant between the residues $i$ and $j$, $\alpha_{ij}$ is the angle between the vector $\hat{R}_j - \hat{R}_i$ and the vector $\Delta \hat{R}_j - \Delta \hat{R}_i$. The parameter $\cos^2(\alpha_{ij})$ is a function of instantaneous conformation, which we approximate by its average value $\langle \cos^2(\alpha_{ij}) \rangle$ and lump into the spring constant term. Also, for simplicity, we assume that spring constants for all interacting pairs are equivalent. Thus, $\Delta U_{ij} = \gamma(\Delta \hat{R}_j - \Delta \hat{R}_i)^2$, where $\gamma$ is the equivalent spring constant for the system. We let this energy to partition equally between the residues $i$ and $j$. Summing up first over all the neighbors of residue $i$ and then over all residues of the protein gives the total energy fluctuation

$$\Delta \hat{U} = \sum_i \Delta \hat{U}_i = \frac{1}{2} \gamma \sum_i \sum_j C_{ij} (\Delta \hat{R}_j - \Delta \hat{R}_i)^2.$$

(8)

The force $F_i$ on the residue $i$ when it is displaced from its mean position by $\Delta \hat{R}_i$ is obtained from the energy as $F_i = \langle \frac{\partial \Delta \hat{U}}{\partial \Delta \hat{R}_i} \rangle = -\gamma \sum_j C_{ij} (\Delta \hat{R}_j - \Delta \hat{R}_i)$ which rearranges into

$$F_i = \Gamma_{ij} \Delta \hat{R}_j,$$

(9)

where the matrix $\Gamma$ is the matrix of the GNM defined as

$$\Gamma_{ij} = \begin{cases} -\gamma C_{ij} & \text{if } i \neq j \\ \gamma \sum_k C_{ik} & \text{if } i = j. \end{cases}$$

(10)

2.6. The heat capacity and fluctuations of energy and entropy

The heat capacity at constant pressure is $C_P = \frac{\partial U}{\partial T} = \frac{\partial U}{\partial T}$, where the second equality follows from the incompressibility assumption of the model. For the canonical ensemble, equation (B.1) gives

$$C_P = \frac{\partial U}{\partial T} = \frac{1}{kT^2} \langle (\Delta \hat{U})^2 \rangle = \frac{1}{kT^2} \left( \sum_i \sum_j \Delta \hat{U}_i \Delta \hat{U}_j \right).$$

(11)

Substitution from equation (8) leads to

$$C_P = \frac{1}{4kT^2} \sum_i \sum_k \sum_j C_{ij} C_{kl} (\Delta \hat{R}_k - \Delta \hat{R}_l)^2 \times (\Delta \hat{R}_k - \Delta \hat{R}_l)^2.$$  

(12)

An equivalent definition of the heat capacity is $C_P = T \frac{\partial S}{\partial T}$. Applying equation (B.1) to the right-hand side of this expression, we obtain

$$C_P = T \frac{\partial S}{\partial T} = \frac{1}{kT} \langle \Delta \hat{S} \Delta \hat{U} \rangle = \frac{1}{kT} \left( \sum_i \sum_j \Delta \hat{S}_i \Delta \hat{U}_j \right).$$

(13)

For the canonical ensemble, unlike the energy and the entropy, the free energy of the system does not fluctuate, as shown in Appendix C. Thus, $\langle (\Delta A)^2 \rangle = 0$, and we obtain $\langle (\Delta \hat{A})^2 \rangle = \langle (\Delta \hat{U} - T \Delta \hat{S})^2 \rangle = 0$. Expanding the right-hand side, and substituting from equations (11) and (13), we obtain the relationship between entropy and energy fluctuations:

$$\langle (\Delta \hat{S})^2 \rangle = T^{-2} \langle (\Delta \hat{U})^2 \rangle.$$  

(14)

Thus, entropy fluctuations are not independent of energy fluctuations.

2.7. Correlations of energy fluctuations

We now examine in more detail the contributions to the heat capacity from interactions of residue pairs:

$$\langle \Delta \hat{U}_i \Delta \hat{U}_k \rangle = \frac{1}{4} \langle kT \rangle^2 \sum_j \sum_l C_{ij} C_{kl} \times (\Delta \hat{R}_k - \Delta \hat{R}_l)^2.$$  

(15)

Expanding the terms in the brackets, and performing the indicated averages as outlined in appendix E, leads to the final expression for correlations of the energy fluctuations:

$$\langle \Delta \hat{U}_i \Delta \hat{U}_k \rangle = \frac{1}{4} \langle kT \rangle^2 \sum_j \sum_l C_{ij} C_{kl} \times \left[ 2(\Gamma_{ik}^{-1})^2 + (\Gamma_{il}^{-1})^2 + (\Gamma_{jk}^{-1})^2 + (\Gamma_{jl}^{-1})^2 \right]$$

$$+ \Gamma_{ii}^{-1} \Gamma_{kk}^{-1} + \Gamma_{ii}^{-1} \Gamma_{ll}^{-1} + \Gamma_{jj}^{-1} \Gamma_{kk}^{-1} + \Gamma_{jj}^{-1} \Gamma_{ll}^{-1} - 4(\Gamma_{il}^{-1} \Gamma_{kl}^{-1} + \Gamma_{ik}^{-1} \Gamma_{jl}^{-1} + \Gamma_{jk}^{-1} \Gamma_{il}^{-1} + 2\Gamma_{ij}^{-1} \Gamma_{jk}^{-1})$$

$$- 2(\Gamma_{ii}^{-1} \Gamma_{kl}^{-1} + \Gamma_{ik}^{-1} \Gamma_{kl}^{-1} + \Gamma_{jk}^{-1} \Gamma_{ii}^{-1} + \Gamma_{il}^{-1} \Gamma_{jk}^{-1})$$

$$+ 4(\Gamma_{ij}^{-1} \Gamma_{kk}^{-1} + \Gamma_{ik}^{-1} \Gamma_{jl}^{-1} + \Gamma_{jk}^{-1} \Gamma_{il}^{-1})).$$

(16)

This is the major result of the present paper.

At a fixed temperature, the correlation given by equation (16) is proportional to the contribution of the interaction of residues $i$ and $j$ to the heat capacity. Summing equation (16) over the index $k$ leads to the energetic interaction, $\Delta U_i$, of residue $i$ with the remaining residues of the protein:

$$\Delta U_i = \sum_k \langle \Delta \hat{U}_i \Delta \hat{U}_k \rangle.$$  

(17)

2.8. The temperature coefficient of the mean-squared distance between two residues

We now investigate the temperature coefficient $\frac{\partial \langle \Delta R_{ij}^2 \rangle}{\partial T}$ of the mean-squared distance between residues $i$ and $j$. This quantity shows the relative amount of the total energy observed by the specific inter-residue interaction. The relationship of the derivative to energy fluctuations follows from equation (B.1) as $\frac{\partial \langle \Delta R_{ij}^2 \rangle}{\partial T} = \frac{1}{kT^2} \langle \Delta \hat{R}_{ij}^2 \Delta \hat{U} \rangle$. In order to further characterize this relation for harmonic interactions, we evaluate the more general term $\langle \Delta R_{ij}^2 \Delta \hat{R}_k^2 \Delta \hat{U} \rangle$ which was previously derived in Halligou et al (2010) and is briefly summarized in appendix F. The resulting expression is

$$\frac{\partial \langle \Delta R_{ij}^2 \rangle}{\partial T} = \frac{1}{kT^2} \langle \Delta \hat{R}_{ij} \Delta \hat{U} \rangle = k(\Gamma_{ii}^{-1} - 2\Gamma_{ij}^{-1} + \Gamma_{jj}^{-1}).$$
Figure 1. Contour diagram for $\langle \Delta \hat{U}_i \Delta \hat{U}_j \rangle$ showing the important correlations between the residues ARG20, TYR21, TYR35 and CYS51.

Summing over the index $j$ in equation (18) leads to the total correlation, $C_{T,i}$, of residue $i$ with all other residues of the protein:

$$C_{T,i} = k \sum_j \left( \Gamma_{ij}^{-1} - 2 \Gamma_{ij}^{-1} + \Gamma_{jj}^{-1} \right).$$

(19)

3. Results and discussion

We now apply the predictions of equations (16) and (17) to the analysis of energy interactions of two widely studied systems: (i) BPTI and (ii) the PDZ domain. In all calculations, the cutoff distance is taken as 7.0 Å. Our interest is in localized motions that identify specific residues. For this reason, we concentrate on the large eigenvalue end of the spectrum of the gamma matrix. We retain the largest five eigenvalues of the gamma matrix in calculating $\Gamma^{-1}$. The eigenvalues beyond the fifth do not contribute much to local events as our calculations show. In previous studies, we concentrated mostly on the largest eigenvalue (Haliloglu and Erman 2009, Haliloglu et al 2010, 2008, Tuzmen and Erman 2011) which accounted for the majority of events associated with the correlations that we studied. Using the five largest eigenvalues now gives a consistent picture of the fine details of the energy–structure relations as we discuss below.

(i) BPTI. Calculations are performed using two pdb structures: 1BPL.pdb and 4PTI.pdb, that are obtained in the presence and absence of the ligand, respectively. The results of equation (16) are presented in a contour diagram in figure 1 for 1BPI. The dark regions show the large values of the correlation $\langle \Delta \hat{U}_i \Delta \hat{U}_j \rangle$ of energy fluctuations between residues $i$ and $j$. According to the results, ARG20 and TYR21 are correlated with TYR35 and CYS51. TYR35 is correlated with CYS51.

Figure 2. Results of calculations based on equation (17) showing the energetic correlation of a given residue. The correlations are given in arbitrary units. The light curve is obtained from the apo form of BPTI, and the solid curve is with the ligand.

The results obtained from equation (17) are shown in figure 2. The thick line is obtained by using the PDB file 1BPI whose crystal structure is obtained in the presence of the phosphate group that binds to ARG20 and TYR35. The two peaks identify the binding site residues for the ligand. The light curve is obtained by using the PDB file 4PTI which is crystallized in the absence of the ligand. The curve obtained is essentially the same as that for the liganded protein. This shows that the information for binding of the ligand at the specified position is already present in the apo form of the protein. CYS30 and CYS51 make a disulfide bridge. Figure 1 shows that there is some but not a strong correlation between these two residues. The two other disulfide bridges, 14–38 and 5–55, do not appear to be interacting energetically according to the present model. Neither of these two is on the interaction pathway of this protein which we define below. PHE45 appears as a small peak in the figure. From figure 1, we see that PHE45 correlates with ARG20, TYR35 and CYS51. PHE45-ARG20 and PHE45-CYS51 are contact interactions, whereas the correlations of PHE45–TYR35 are long-distance correlations as seen from figure 3.
Figure 4. The six peaks of energy correlation determining the interaction path in the PDZ domain for the protein 3I4W.

Figures 3(a) and (b) show the interaction path that is obtained with the present study. The ligand PO4 is shown in red. It makes a contact with ARG20 and TYR35 as shown in figure 3(b). The path through which energy is transmitted in the protein is summarized by figure 3(b). The path starts with ARG20 and TYR35, goes through TYR21 which makes a distance of 2.8 Å with PHE45, which in turn makes a distance of 3.96 Å with CYS51. Finally, CYS51 makes a sulfide bridge with CYS30. Thus, the calculations show that the interaction pathway in BPTI is through ARG20-TYR21-PHE45-CYS51-CYS30. TYR35 and ARG20 form the energy gate at one end of the pathway, and CYS30 is at the other end.

(ii) The PDZ system. The second application of the energy fluctuation model is on a set of 12 proteins with the PDZ domain. Their Protein Data Bank identities are 3I4W, 3NGH, 3JXT, 3QIK, 2KQF, 2W7R, 3PS4, 3QJM, 2KAW, 2KOJ, 3KHF, 2KG2. The correlations of energy fluctuations are calculated from equation (17) and the results are presented in the 12 panels of figure 6. All of the proteins in this group exhibit six characteristic peaks that are similarly located on the primary sequence of each protein. We consider 3I4W in detail here. In figure 4, results of calculations based on equation (17) are presented, where energy correlations of residues identified by the residue index along the abscissa are presented in arbitrary units. For uniformity, residue indices are numbered from 1 to N and do not correspond to the actual residue numbers given in the data bank files. Six major peaks are observed, numbered from left to right in the figure. The residues corresponding to the six peaks are ILE316 at peak 1, PHE325, ASN326 and ILE327 at peak 2, ILE336 at peak 3, PRO346 at peak 4, ASP357 at peak 5, and GLN391 at peak 6. The corresponding structures are shown in figure 5 in 3D.

In figure 6, the energy correlation peaks are shown for all of the 12 proteins that we investigated. In all of the proteins, the characteristic six peaks are observable. In some cases, there are shifts in the locations of the peaks and in their amplitudes due to differences in the numbering of the residues and due to effects from the diverse neighborhoods of the proteins, but in all cases the characteristic peak structure is recognizable and is in general agreement with the patterns suggested for similar systems by Lockless and Ranganathan (1999).

The primary aim of this paper was to establish a connection between a statistical mechanical description of proteins and energetic interaction that define their function. The structural basis of energetic perturbations and fluctuations entered into the model in a simple way through the connectivity matrix, thus establishing the relation of processes to protein architecture. In this sense, this study outlines the conceptual framework for the energy processes in proteins. It can be improved in several directions. Although a harmonic potential is used for the examples, the general thermostatistics treatment is not confined to harmonic interactions. Anharmonicities may be introduced by the use of equation (D.1) by a suitable choice of force-fluctuation equations of state. The two examples presented above are intended for a proof of principle. More detailed analysis is needed to demonstrate the capabilities of the model. Our recent work (Tuzmen and Erman 2011) on a related formulation, applied to 24 benchmark proteins, is encouraging in this respect. The cutoff distance of
7.0 Å seems to be arbitrary, as has been the case in all previous studies of the GNM. A scaling study of the cutoff distance based on 4810 non-redundant structures obtained from http://dunbrack.fccc.edu/Guoli/culledpdb suggested that 7.0 Å is a reasonable value for the cutoff distance. Keeping only the five largest eigenvalues in the formulation to represent localized events also seems arbitrary, and further work on this is needed. For the present paper, five eigenvalues were necessary and sufficient to reflect the basic features of the PDZ domain, for example. Fewer eigenvalues reflected some but not all features, and a larger number of eigenvalues added to the redundancy in the results. The energy interaction pathways presented in figures 3 and 5 do not lie on a straight line but are rather of fractal nature. There should be a deep relation between the fractal nature of the pathways and the protein function, which we cannot see now, but is definitely worthy of further examination. Finally, it seems possible to apply the model to the determination of unknown domains of interaction in a diverse set of proteins, simply by searching similar peaks, as it is the case with the PDZ domain.

Acknowledgments

Support by the Turkish Academy of Sciences is gratefully acknowledged.

Appendices

Appendix A. The probability distribution of fluctuations (from Callen (1985))

The general form of the probability function for the instantaneous values $\hat{X}$ of the thermodynamic variables is

$$f(\hat{X}) = \exp \left\{ -\frac{1}{k} \left[ F_0, ..., F_m \right] - \frac{1}{k} \sum F_i \hat{X}_i \right\}. \quad \text{(A.1)}$$

Here, $\hat{X}_i$ are the instantaneous values of the parameters of the model, which in general could be the energy and position of each residue, their volumes, electric field acting on each residue, etc.; $F_i$ are the corresponding entropic variables, such as

![Figure 6. The energy correlation diagrams for the 12 proteins that contain the PDZ domain.](image)
as \( \frac{1}{r}, -\frac{\partial}{\partial r}, \frac{\partial}{\partial \phi}, -\frac{\partial}{\partial \phi} \), etc showing the constant force acting on residue \( i \), with \( P \) the pressure and \( E_i \) the electric field on residue \( i \). \( S[F_0, ..., F_m] \) is the general Massieu transform of the entropy with respect to its arguments. In this paper, we assume that the protein exchanges energy with its surroundings only, no constant forces and electric field acts on it, and therefore \( F_0 = \frac{1}{r}, \dot{\phi} = -\frac{\partial}{\partial r}, \) and \( S\left[\frac{1}{r}\right] = S - \frac{1}{r} U \). The resulting canonical form is equation (5) in the text.

**Appendix B. Temperature derivative of a thermodynamic variable (from Prabhu and Sharp (2005))**

For the canonical ensemble, using the probability function given by equation (5), the following relation is obtained:

\[
\frac{d(\dot{X})}{dT} = \frac{1}{kT^2} (\dot{X} \dot{U}) - \langle \dot{X} \rangle \langle \dot{U} \rangle = \frac{1}{kT^2} (\langle \dot{X} \dot{U} \rangle).
\]  

**(B.1)**

**Appendix C. The expression \( \langle (\Delta A)^2 \rangle = 0 \) (from Meirovitch (1999))**

Writing the Helmholtz free energy, \( A \), as \( A = \sum_i f_i A_i = \sum_i f_i (U_i + kT \ln f_i) \) and substituting for the probability from equation (5), i.e. \( f_i = \exp \left[ -\frac{1}{kT} \dot{U} \right] / Z \), gives \( A_i = -kT \ln Z \), which is independent of the instantaneous states of the protein, hence does not fluctuate.

**Appendix D. Fluctuation recursion relation (from Callen (1985))**

For any general system, the correlation of the fluctuations of any two thermodynamic variables \( \Delta X_j \) and \( \Delta X_k \) is given by the derivative

\[
\langle \Delta X_j \Delta X_k \rangle = -k \left( \frac{\partial X_j}{\partial F_k} \right)_{F_1, ..., F_{k-1}, F_{k+1}, ...}
= kT \left( \frac{\partial X_j}{\partial F_k} \right)_{F_1, ..., F_{k-1}, F_{k+1}, ...}.
\]

**(D.1)**

Repeated use of this expression for any fluctuating function \( \phi \) gives

\[
\langle \phi \Delta X_j \Delta X_k \rangle = -k \left( \frac{\partial \phi}{\partial F_k} \right) = -k \frac{\partial \phi}{\partial F_k} + k \left( \frac{\partial \phi}{\partial F_k} \right).
\]

The thermostatistical basis of the GN1 equation

\[
\langle \Delta R \Delta R \rangle = kT (\Gamma^{-1})_{jk}
\]

is obtained from equation (D.1), by choosing \( F_i = \Gamma_{ij} \Delta R_j \)

**Appendix E. Fourth-order correlations of residue fluctuations**

Expanding the right-hand side of equation (16) leads to

\[
\langle \Delta U_i \Delta U_i \rangle = \frac{1}{4} \gamma^2 \sum_i \sum_j C_{ij} C_{kl} \left[ \langle \Delta R_i^2 \Delta R_i^2 \rangle + \langle \Delta R_i^2 \Delta R_j^2 \rangle \right]
+ \langle \Delta R_i^2 \Delta R_i^2 \rangle + \langle \Delta R_i^2 \Delta R_i^2 \rangle - 2 \left( \langle \Delta R_i^2 \Delta R_i \Delta R_j \Delta R_k \rangle + \langle \Delta R_i^2 \Delta R_i \Delta R_j \Delta R_k \rangle \right)
+ \langle \Delta R_i^2 \Delta R_i \Delta R_j \Delta R_k \rangle + 4 \langle \Delta R_i \Delta R_j \Delta R_k \Delta R_l \rangle.
\]

**(E.1)**

Using relations (D.1) and (D.2), and the relation \( F_i = \Gamma_{ij} \Delta R_j \), the higher moments of fluctuations are replaced by the products of the matrix \( \Gamma \) as follows:

\[
\langle \Delta R_i^2 \Delta R_i^2 \rangle = \left[ 2 \langle \Gamma_{ij}^{-1} \rangle + \langle \Gamma_{ij}^{-1} \Gamma_{kl} \rangle \right] \frac{(kT)^2}{2}
\]

\[
\langle \Delta R_i^2 \Delta R_i^2 \rangle = \left[ 2 \langle \Gamma_{ij}^{-1} \rangle + \langle \Gamma_{ij}^{-1} \Gamma_{kl} \rangle \right] \frac{(kT)^2}{2}
\]

\[
\langle \Delta R_i^2 \Delta R_i^2 \rangle = \left[ 2 \langle \Gamma_{ij} \Gamma_{kl} \rangle - \langle \Gamma_{ij} \rangle \langle \Gamma_{kl} \rangle \right] \frac{(kT)^2}{2}
\]

\[
\langle \Delta R_i^2 \Delta R_i^2 \rangle = \left[ 2 \langle \Gamma_{ij} \Gamma_{kl} \rangle - \langle \Gamma_{ij} \rangle \langle \Gamma_{kl} \rangle \right] \frac{(kT)^2}{2}
\]

\[
\langle \Delta R_i^2 \Delta R_i^2 \rangle = \left[ 2 \langle \Gamma_{ij} \Gamma_{kl} \rangle - \langle \Gamma_{ij} \rangle \langle \Gamma_{kl} \rangle \right] \frac{(kT)^2}{2}
\]

**(F.2)**

Starting with the definition of energy fluctuation (Haliloglu et al 2010, Tuzmen and Erman 2011)

\[
\Delta U = F_i \Delta R_i = F_i \langle \Gamma^{-1} \rangle_{jk}
\]

and applying expressions (D.1) and (D.2), we obtain

\[
\langle \Delta U \Delta R_i \Delta R_j \rangle = kT \langle \Delta R_i \Delta R_j \rangle = (kT)^2 (\Gamma^{-1})_{jk}.
\]

**(F.2)**

Using the relation \( \langle \Delta R_i \rangle = kT \langle \Gamma_i^{-1} \rangle - 2 \langle \Gamma_{ij}^{-1} \rangle \) and (F.2) leads to equation (18).

**References**

Bahar I, Atilgan A R and Erman B 1997 Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential Folding Des. 2 173–81

Callen H B 1985 Thermodynamics and an Introduction to Thermostatistics (New York: Wiley)

Cooper A 1976 Thermodynamic fluctuations in protein molecules Proc. Natl Acad. Sci. 73 2740–1

Fenimore P W, Frauenfelder H, McMahon B H and Parak F G 2002 Slaving: solvent fluctuations dominate protein dynamics and functions Proc. Natl Acad. Sci., USA 99 16047–51

Ferrii E, Pasta J and Ulam S 1955 Studies of nonlinear problems Los Alamos Rep. LA-1940 pp 977–88

Frauenfelder H, Fenimore P W and McMahon B H 2002 Hydration, slaving and protein function Biophys. Chem. 98 35–48
Freire E 1999 The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme Proc. Natl Acad. Sci., USA 96 10118–22
Gur M and Erman B 2010 Quasi-harmonic analysis of mode coupling in fluctuating native proteins Phys. Biol. 7 046006
Halliloglu T and Erman B 2009 Analysis of correlations between energy and residue fluctuations in native proteins and determination of specific sites for binding Phys. Rev. Lett. 102 088103
Halliloglu T, Gul A and Erman B 2010 Predicting important residues and interaction pathways in proteins using Gaussian network model: binding and stability of HLA proteins Plos Comput. Biol. 6 e1000845
Halliloglu T, Seyrek E and Erman B 2008 Prediction of binding sites in receptor–ligand complexes with the Gaussian network model Phys. Rev. Lett. 100 228102
Hilser V J, Dowdy D, Oas T G and Freire E 1998 The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble Proc. Natl Acad. Sci., USA 95 9903–8
Hinsen K 1998 Analysis of domain motions by approximate normal mode calculations Proteins: Struct. Funct. Genet. 33 417–29 (http://dunbrack.fccc.edu/Guoli/culledpdb Accessed 2011)
Juanico B, Sanejouand Y H, Piazza F and De Los Rios P 2007 Discrete breathers in nonlinear network models of proteins Phys. Rev. Lett. 99 238104
Kabakcioglu A, Yuret D, Gur M and Erman B 2010 Anharmonicity, mode-coupling and entropy in a fluctuating native protein Phys. Biol. 7 046005
Kloczkowski A, Mark J E and Erman B 1989 Chain dimensions and fluctuations in random elastomeric networks: I. Phantom Gaussian networks in the undeformed state Macromolecules 22 1423–32
Leitner D M 2008 Energy flow in proteins Annu. Rev. Phys. Chem. 59 233–39
Lockless S W and Ranganathan R 1999 Evolutionarily conserved pathways of energetic connectivity in protein families Science 286 295–9
Meirovitch H 1999 Simulation of a free energy upper bound, based on the anticorrelation between an approximate free energy functional and its fluctuation J. Chem. Phys. 111 7215–24
Monod J, Wyman J and P C J 1965 On the nature of allosteric transitions: a plausible model J. Mol. Biol. 12 98–118
Moritsugu K, Miyashita O and Kidera A 2000 Vibrational energy transfer in a protein molecule Phys. Rev. Lett. 85 3970–3
Oylumluoglu G, Buyukkilic F and Demirhan D 2006 Investigation of hydration effect of the proteins by phenomenological thermostatistical methods Physica A 361 255–62
Oylumluoglu G, Buyukkilic F and Demirhan D 2007 Investigation of heat capacities of proteins by statistical mechanical methods Physica A 375 577–83
Piazza F, De los Rios P and Sanejouand Y H 2005 Slow energy relaxation of macromolecules and nanoclusters in solution Phys. Rev. Lett. 94 145502
Piazza F and Sanejouand Y H 2008 Discrete breathers in protein structures Phys. Biol. 5 026001
Piazza F and Sanejouand Y H 2009 Long-range energy transfer in proteins Phys. Biol. 6 046014
Prabhu N V and Sharp K A 2005 Heat capacity in proteins Annu. Rev. Phys. Chem. 56 521–458
Rader A J 2010 Coarse-grained models: getting more with lessCurr. Opin. Pharmacol. 10 753–9
Sturtevant J M 1977 Heat capacity and entropy changes in processes involving proteins Proc. Natl Acad. Sci. 74 2236–40
Suel G M, Lockless S W, Wall M A and Ranganathan R 2003 Evolutionarily conserved networks of residues mediate allosteric communication in proteins Nat. Struct. Biol. 10 59–69
Swain J F and Gierasch L M 2006 The changing landscape of protein allostery Curr. Opin. Struct. Biol. 16 102–8
Tuzmen C and Erman B 2011 Identification of ligand binding sites of proteins using the Gaussian network model Plos One 6 e16474
Yogurtcu O N, Gur M and Erman B 2009 Statistical thermodynamics of residue fluctuations in native proteins J. Chem. Phys. 130 095103