Network model of a protein globule

E. Z. Meilikhov · R. M. Farzetdinova

Received: 8 February 2013 / Accepted: 24 June 2013 / Published online: 30 July 2013
© Springer Science+Business Media Dordrecht 2013

Abstract Phase transition of a protein globule is considered in the frameworks of (i) the generalized mean-field theory for the order parameter, characterizing the extent of the deviation of a protein three-dimensional structure from its native state and (ii) the network model that treats a protein globule as a small-world network with a significant percent of long-range links between amino acid residues. Temperature dependencies of the introduced order parameter are defined and phase-transition temperatures are found on the basis of the function defining the distribution of links' numbers for protein residues. An important role of long-range links, promoting considerable rise of thermal protein stability, is demonstrated by the example of a correlation between protein melting temperature and a fraction of disulfide bonds.

Keywords Protein globule · Network model · Small-world network · Phase transition · Order parameter · Mean-field theory · Long-range links

1 Introduction

In the native state, proteins (or at least a significant amount of them) are not quasi-one-dimensional objects like beads, consisting of amino acid residues with strong polypeptide bonds but three-dimensional objects—globules that originate due to folding provoked by the weak interaction of residues with each other and solvent molecules. The structural properties of these objects could be described in terms of the graph (network) theory, which describes (if we abstract from the specific nature of interacting elements) universal features of these systems, allows to classify them according to some indicants, to estimate their stability qualitatively etc. [1–3]. In the context of this theory, this principally new quality, which a protein acquires after its folding into the compact globule, consists of long links appearing between residues before being well spaced along the original protein chain. In

E. Z. Meilikhov (✉) · R. M. Farzetdinova
Kurchatov Institute, 123182 Moscow, Russia
e-mail: meilikhov@yandex.ru

the denatured quasi-one-dimensional state, there are no such long links. This new feature allows to refer the network of protein inter-residue links to so-called small-world networks (see below).

Currently, graph theory finds an ever-growing use for describing properties of biological systems [1]. Formally, it is convenient to use the so-called connectivity matrix $\mathbb{A}$ generated by a simple recipe—its elements A_{ij} are different from zero (and equal to unity) only if the i -th and j -th chain residues ($i \neq j$) are connected to each other:

$$A_{ij} = \begin{cases} 1 & (i \text{ and } j \text{ are connected, } i \neq j) \\ 0 & (i \text{ and } j \text{ are not connected}) \end{cases}. \quad (1)$$

The criterion of residues' connectivity can be chosen in a number of ways. Usually, residues are considered as being linked if they are removed in real space (not along the chain!) at a distance less than some critical one: $r \leq r_c$. The latter is determined by the weak interaction radius and supposed to be equal to $r_c = 7\text{--}8 \text{ \AA}$ [4, 5].

This matrix fully describes all characteristics of a network containing N nodes: (i) local parameters

- *the mean connectiveness* (determines the mean number of links falling on a node)

$k = \frac{1}{N} \sum_{i=1}^N k_i$, where $k_i = \sum_{j=1}^N A_{ij}$ is the number of links for the i -th node,

- *clustering coefficient* (determines the probability that neighbors of the i -th node are

also neighbors of each other) $C = \frac{1}{N} \sum_{i=1}^N C_i$ where $C_i = \frac{\sum_{j=1}^N \sum_{l=1}^N A_{ij} A_{il} A_{jl}}{\sum_{j=1}^N \sum_{l=1}^N A_{ij} A_{il}}$ is the clustering coefficient of the i -th node, and (ii) the global network parameter

- *the mean effective distance* between nodes

$$L = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N L_{ij},$$

where $L_{ij} = \begin{cases} |i-j|, & (A_{ij} = 0) \\ 1, & (A_{ij} = 1) \end{cases}$ is the distance between nodes (i, j) .

Two limiting cases of the network structure are a regular and random one [2]. The first case could be exemplified by a volleyball net consisting in rectangle cells being repeated regularly. For regular networks

$$L_{\text{reg}} \sim N, \quad C_{\text{reg}} \sim 1.$$

The random network results from a system of originally isolated nodes if any two of them are connected to each other with some probability. For random networks

$$L_{\text{rand}} \sim \ln N, \quad C_{\text{rand}} \propto 1/N.$$

However, it turns out that for a number of real networks (transport, social, Internet, etc.) and, among them, for the network of interacting residues in a protein globule [6–8]

$$L \sim L_{\text{rand}} \sim \ln N, \quad C \sim C_{\text{reg}} \sim 1,$$

that is the mean distance between their nodes is much smaller than in the regular network, while high clustering is preserved. Such networks have derived their name from the proverb

“It’s a small world!” [9] because a significant amount of long links provide connectivity between a pair of any nodes via small amounts of intermediary nodes. This phenomenon is known in sociology as the six degrees of separation rule [10].

Investigating connectivity matrices for various proteins of known structures with residues’ number $N = 10^2\text{--}10^3$ has led to the understanding that in the native state they represent small-world networks [7]: $L \approx 4 \sim \ln N$, $C \approx 0.6 \sim 1$. The same result has been demonstrated in [6]: $L \sim L_{\text{rand}}$, $C \gg C_{\text{rand}}$. Small-world networks exhibit enhanced stability because breaking some of their links does not destroy network functionality.

The qualitative characteristics of this stability could be defined in a variety of ways. For instance, in [11] the network strength has been specified through the *weighted* connectivity matrix

$$A_{ij}^w = \begin{cases} A_{ij} w_{ij} & (i \text{ and } j \text{ are linked, } i \neq j \pm 1) \\ 2.55 & (i = j \pm 1), \end{cases} \quad (2)$$

where in the elements $A_{i, i \pm 1} = 2.55$ correspond to the mean absolute energy of strong covalent connections (in RT units) and w_{ij} are the weights of the weak links between residues i and j being numerically equal to (taking with the opposite sign) energies of those links (in RT units) presented in [4]. The network strength $S = \sum_i S_i$ is the sum of strengths $S_i = \sum_j A_{ij}^w$ of its various nodes. Being multiplied by the properly weighted network connectivity, it provides the scoring function that predicts the result of interacting proteins with ligands successfully. Notice, that the contribution of covalent links in (2) influences only the background against which the folding occurs.

The folding rate could also be predicted proceeding from characteristics of the network corresponding to linked protein residues. In the paper [12], this rate (more exactly, its logarithm) has been shown to be well correlated with the relative contact order

$$CO = \frac{1}{\eta N} \sum_{i,j} S_{ij}, \quad (3)$$

where S_{ij} is the distance (along the chain) between *interacting* residues i, j , or the number of residues between them and η is the number of *interacting* residue pairs in a protein. Long-range links between residues, to which large S_{ij} values correspond, make easier the process of protein folding but increasing the number η of interacting residues enhances this process as well. Thus, the absolute contact order $Abc_CO = \eta \cdot CO$ (introduced in [13]) seems intuitively to be more appropriate for estimating the folding rate. In fact, it turns out to be better correlated with the latter [14].

Algorithms of transforming a regular network into a small-world one are known [15]: (i) in the original regular network, some fraction p of links is randomly chosen and their ends are disconnected from nodes, (ii) one of the ends of each of these links is connected to some randomly chosen network node and, finally, (iii) the second (left free) end of the link is connected to a randomly chosen node without forming a loop (which is the link whose both ends connect with the same node) and doubling existing links. It is clear that due to the random selection of nodes, the great majority of newly generated links are long-range originally missing ones whose lengths are much longer than the period of the original structure.

Two different approaches are possible to describe properties of proteins considered as networks of amino acid residues interacting with each other (see below). Within the framework of one of them, a protein is a *random* network structure whose description is possible on the basis of the properly generalized mean-field theory. Within the other

approach, a protein is the *small-world* network with a significant fraction p of long-range links sensibly influencing its stability. The following consideration refers mainly to moderate (single-domain) water-soluble globular proteins.

2 Materials and methods

To consider qualitatively the transformation process *native protein globule* $\rightleftharpoons$ *coil* (*melted globule*), being the phase transition [16], one needs to introduce the order parameter—thermodynamic value characterizing the long-range order in a system [17]. The steady-state order parameter equals zero in the disordered phase and nonzero in the ordered one. To define such a parameter for the protein network, one could use the analogy of magnetic ordering in a system of Ising magnetic moments having two possible orientations only—parallel or anti-parallel to some appropriate direction (up or down). The order parameter in such a system is defined as follows [18]. If out of the total number N of magnetic moments N^+ moments are oriented up and $N^- = N - N^+$ —down, then the total magnetic moment M of the system is proportional to the difference $N^+ - N^-$. Let us define the dimensionless relative magnetization j as the ratio M/M_s of the total magnetic moment M of the system and its maximum magnetic moment M_s in the saturated state, where $N^+ = N$ (in this case all magnetic moments direct upwards and, hence, $M_s \propto N$). Evidently, the relative magnetization equals $j = (N^+ - N^-) / N = 2N^+ / N - 1$ and its absolute value is within the limits $0 \leq j \leq 1$. At $j = 1$ ($N^+ = N$), the magnetic system is completely ordered and at $j = 0$ ($N^+ = N^-$) it is fully disordered. The magnetization j is just playing the role of the order parameter in the system considered. Fractions N^+ / N and N^- / N of the up and down states are connected with j through the simple relation $N^\pm / N = (1 \pm j) / 2$. Similarly, one could use the analogy of melting molecular liquid crystals, which is described through the use of the parameter of orientational order $q = 2(N^+ / N - 1)$, where N^+ is the fraction of molecules oriented in the right direction [19].

The translation of these definitions into the system of amino-acid protein residues is straightforward. Specifically, let N be the total number of residues in the protein chain and only N^+ of them are in equilibrium states corresponding to the native conformation while, due to thermal fluctuations, the rest of the N^- residues are in states that are far from equilibrium. The degree of this ‘farness’ is defined not by absolute values of space and/or angle deviations of residues from the equilibrium state but by the condition that in the perturbed state energies of weak residues’ interaction are significantly (but still by a value less than the thermal energy, see below) different from equilibrium energies. Then the order parameter ε is defined by the analogy of the above:

$$\varepsilon = \frac{N^+ - N^-}{N}, \quad \frac{N^\pm}{N} = \frac{1 \pm \varepsilon}{2}. \quad (4)$$

The order parameter, introduced in such a way, is close to the fraction of monomer units connected with hydrogen bonds that has been estimated in a number of experiments [20]. Our choice for this parameter is associated unambiguously with the fraction of the right (native) links between residues and gives the convenient diapason $[0, 1]$ of the order parameter variation. This simple approach is similar to that considered in papers [21, 22] whose authors declared: “The nonnative interactions of adjoining residues will have a distribution of energies with mean $-J''$. The native interactions must satisfy the principle of minimal frustration and the simplest way to ensure this is to set all native tertiary interaction energies equal to $-J'$ where $J' > J''$ ”. Obviously, that approach is equivalent to ours.

Though the interaction energy of various residue pairs is different [4], they are all of the same order (by interaction energy it is here meant one that is renormed by interaction with the solvent). Thereupon, remaining in the framework of the semiquantitative model we will further consider all those energies being identical and equal to J , so that the interaction energy W_{nm} of any two neighbor residues (in equilibrium and perturbed states) could be written in the form typical for an Ising system:

$$W_{nm} = -Js_n s_m \quad (n, m = 1, 2, \dots, N; n \neq m), \quad (5)$$

where $s_{n,m} = 1$ and $s_{n,m} = -1$ for residues being in native and perturbed states, correspondingly (these parameters are analogous to Ising model spins). The total interaction energy W_n of the n -th residue with all its neighbors (naturally, excluding its strong covalent binding in the protein chain) is

$$W_n = \sum_{\text{neighbors}} W_{nm} = -Js_n \sum_{\text{neighbors}} s_m, \quad (6)$$

where the summation is performed over all those neighbors in the protein globule.

In the standard mean-field theory, the number of neighbors z is the same for all network nodes and the last sum in (6) is assumed to be equal to its mean value $\langle \sum s_m \rangle = z\varepsilon$, so that the mean interaction energy (per unit residue) is $\langle W_n \rangle = -Jz\varepsilon$. The respective standard mean-field equation determining the order parameter ε as a function of temperature T reads [18]

$$\varepsilon = \tanh \left(\frac{zJ}{k_B T} \varepsilon \right) \quad (7)$$

(k_B is the Boltzmann constant). In particular, it also predicts the temperature $T_c = zJ$ of the phase transition.

By field it is here meant some effective field h , exerted on a given network node (amino-acid residue, in our case) by all its neighbors. Quantitatively, this field is proportional to the order parameter ε and could be expressed in terms of the mean interaction energy: $h = -\langle W_n \rangle = \lambda\varepsilon$, where $\lambda = zJ$ is the so-called mean-field constant. In the traditional theory, the equivalence of all network nodes is assumed. This is due to the fact that the effective field (that is, the interaction energy W_n) is the same in all nodes, though, in fact, it changes from node to node randomly and one needs to take into account this randomness (at least, approximately). To do this, standard equation (7) could be replaced with its generalized analog [23]

$$\varepsilon = \int_0^\infty \tanh \left(\frac{W}{k_B T} \right) F(\varepsilon; W) dW, \quad (8)$$

where $F(\varepsilon; W)$ is the distribution function of residues' interaction energies for the system with a given order parameter ε . Then the problem is reduced to determining this distribution function for the system considered (the protein globule).

For this purpose, it is particularly necessary to use the distribution functions $f(n)$ of link numbers that are known (based on the information about the spatial proteins structure, see, for instance, Berman et al. [24]) for a number of proteins [4–6, 16, 25]. These functions

are more or less universal for different residues and proteins and their envelopes are well described by the Gauss formulae

$$f(n) = \frac{\exp[-(n - n_0)^2/2\sigma^2]}{\sum_{n=1}^{\infty} \exp[-(n - n_0)^2/2\sigma^2]} \quad (n = 1, 2, \dots), \quad (9)$$

where the most probable number n_0 of links depends on the residue type (n_0 is minimum for Gln, Ser, Asn, Asp, Glu, Lys, and maximum for Leu, Phe, Ile, Met, Val) and the dispersion σ of the link numbers is practically the same for all residues [4]. It is natural to suggest that proteins, consisting in a larger number of residues with a small number of links, are more prone to be disordered wholly or in part. This is confirmed by available data [26].

To simplify further, we neglect the n_0 -difference for various residues and assume $n_0 = 7$, $\sigma = 2$ [6]. Notice, however, that as expected, the average links' number for residues, situated in the globule core, is appreciably larger than that for residues, placed in the near-surface layer [6]. This leads to lesser stability of this layer against thermal fluctuations. Interestingly also, links between residues being distant from each other along the protein chain (by more than 10 intermediate links) represent a small but noticeable fraction of all links (~ 5 – 10% [5]), which is typical for the small-world network (see above). Taking account of these details allows to refine the model.

Let us now define what could be the set of possible energies of an interaction of a given residue with its neighbors in a protein network if this residue has n links with its neighbors and each of them provides equal interaction energy $\pm J$ (the sign depends on the mutual orientations of the interacting residues, see (5)). For certainty, let the spin of a given residue equal $s_n = +1$. Then the probability $p_n(k)$ of k out of n neighbor residues having spins of opposite sign equals

$$p_n(k) = C_n^k \left(\frac{N^+}{N} \right)^{n-k} \left(\frac{N^-}{N} \right)^k = C_n^k \left(\frac{1+\varepsilon}{2} \right)^{n-k} \left(\frac{1-\varepsilon}{2} \right)^k = \frac{1}{2^n} C_n^k (1+\varepsilon)^{n-k} (1-\varepsilon)^k, \quad (10)$$

where C_n^k is the binomial coefficient defining the number of combinations of n elements taken k at a time. For this link configuration, the energy of the residue interaction with its neighbors is equal to $W_n(k) = (n - 2k)J$. Taking into account that the probability to have just n links for a given residue equals $f(n)$, one finds the distribution function of interaction energies for amino acid residues possessing n links:

$$F_n(\varepsilon; W) = \frac{1}{2^n} f(n) \sum_{k=0}^n C_n^k (1+\varepsilon)^{n-k} (1-\varepsilon)^k \delta[W - (n - 2k)J], \quad (11)$$

where $\delta[W - (n - 2k)J]$ is the delta function.

Finally, the total distribution function of residue interaction energies in the system, where they have different numbers of links with their neighbors ($n = 1, 2, 3, \dots$), is obtained by summation of partial (for a given n -value) distribution functions (11):

$$F(\varepsilon; W) = \sum_{n=1}^{\infty} F_n(\varepsilon; W) = \sum_{n=1}^{\infty} \frac{1}{2^n} f(n) \sum_{k=0}^n C_n^k (1+\varepsilon)^{n-k} (1-\varepsilon)^k \delta[W - (n - 2k)J] \quad (12)$$

(The lower limit of the summation over n is chosen to be $n = 1$, because at $n = 0$ the interaction energy is zero and such a residue, without links, does not influence the order parameter of the system.)

The generalized mean-field equation, defining the order parameter in the considered system, is obtained by substituting the distribution function (12) into (8):

$$\varepsilon = \sum_{n=1}^{\infty} \frac{1}{2^n} f(n) \sum_{k=0}^n C_n^k (1 + \varepsilon)^{n-k} (1 - \varepsilon)^k \tanh \left[\frac{(n - 2k)J}{k_B T} \right]. \quad (13)$$

The right-hand side of this equation could be shown to be polynomial containing odd powers of ε only. Hence, its solution uniquely determines only the squared-order parameter: $\varepsilon^2 = [\varepsilon(n_0, \sigma, J/k_B T)]^2$. In our model, negative values of ε do not have physical sense and reflect merely the convention of the up and down concept (in contrast to the magnetization problem for Ising spins, where the j -sign defines the real direction of the magnetic moment arising due to the spontaneous breaking of the symmetry).

3 Results and discussion

Equation (13) defines the temperature dependence $\varepsilon(T)$ of the order parameter and the mean-field approximation describes the conformational transition *protein globule - core* (melted globule) with varying temperature. In addition, one could consider a protein in solvents of various pH. The protein charge depends on the pH and increases at the deviation from the isoelectric point in either direction. This results in mutual repulsion of parts within the protein [27, 28], which is equivalent to the renormalization $J = J(\text{pH})$ of the interaction energy by a solvent and could, in principle, give the pH-dependence $T_c(\text{pH})$ of the phase transition temperature as well. Notice, that the energy J includes, along with the interaction energy, some entropy contribution so that in general $J < W_{nm}$.

In Fig. 1 temperature dependencies $\varepsilon(T)$ of the order parameter are shown for various values of $n_0 = 5-8$ and $\sigma = 1-3$. The average number of the inter-residue links provides the phase transition temperature (the larger the n_0 the higher the latter) and the dispersion σ of

Fig. 1 Temperature dependencies of the order parameter $\varepsilon(T)$ for various average numbers n_0 of weak links per one residue and widths σ of the distribution function of links' numbers (see (16))

the links' numbers distribution function defines the transition width (increasing σ results in smearing the transition). Comparing experimental values of the characteristic temperature and width for the transition *protein globule - core*, one could determine average values of parameters J , σ . As can be seen from Fig. 1, near the transition $kT/J = 4-6$, which gives $J \approx 0.1$ cal/mol at $T \approx 300$ K, there are 1–2 orders of value less than typical values of hydrogen and dipole–dipole bonds. This agrees with protein functionality requiring some flexibility of its structure [29]. Experimental transition width $\Delta T \approx 15-50$ K leads to $\sigma \sim 1$.

The model considered could be refined taking into account that proteins in native states are non-uniform: it has a hydrophobic core that is enveloped by the layer of hydrophilic surface residues of thickness $\Delta R \sim 4$ Å [6]. The topological properties of these two regions are different: although distributions over residues' link numbers are similar to Gaussian distributions (9) of width $\sigma \approx 2$, the mean links' number $n_{01} \approx 8$ in the core is significantly larger than in the envelope, where $n_{02} \approx 5$ [6]. This naturally yields that with heating, the envelope is melted at first and the core is destroyed afterwards [16]. The same phenomenon is responsible for decreasing the melting temperature of nanoparticles [30]. In respect to proteins, this question has been closely discussed in [31] by means of the Lindemann melting criterion. The authors stated that native proteins are surface-molten solids and concluded that the existence of a solid-like core complete with liquid-like motions in surface regions is necessary for protein activity.

At a first approximation, one could consider these regions as independent, introduce two own order parameters ε_1 and ε_2 being, as before, defined by (13) and examine their temperature dependencies separately. In calculations, one should take into account that mean energies of interaction between core residues J_1 and shell residues J_2 could be different. Then the average order parameter defining, for example, microcalorimetric results, could be written as follows

$$\varepsilon(J_1, J_2, T) = v_1 \varepsilon_1(J_1, T) + v_2 \varepsilon_2(J_2, T) = (1 - v_2) \varepsilon_1(J_1, T) + v_2 \varepsilon_2(J_2, T), \quad (14)$$

where v_1 , v_2 are volume fractions of the core and shell, respectively. For example, for a globule of radius $R = 3$ nm, one finds $v_2 \approx 3\Delta R/R \approx 0.3$. The interaction energy of hydrophilic residues situated near the surface of the protein globule is, on average, higher than that of hydrophobic residues composing its core: $J_2/kT \approx 0.2-0.3$, $J_1/kT \approx 0.1-0.2$ [4].

In Fig. 2 there are temperature dependencies $\varepsilon(T)$ found in such a way for the system that consists in two independent regions (core and shell) with the relevant volume fractions $v_1 = 0.7$, $v_2 = 0.3$ and Gaussian distribution functions of links' numbers identified by the parameters $n_{01} = 8.5$, $n_{02} = 5$, $\sigma = 2$. Calculations have been performed for two cases: $J_2/J_1 = 1$ and $J_2/J_1 = 1.5$. The above introduced order parameter ε is some an indirect measure of the system energy. Therefore, peculiarities of its temperature derivative $d\varepsilon/dT$ are associated with those of the system's heat capacity. The latter are conveniently studied by means of calorimetry [28]. In the insert of Fig. 2, there are the temperature dependencies of this derivative simulating an external view of experimental curves, obtained with the microcalorimetry method in the course of thermal protein denaturation [32].

Let us now proceed to the quantitative estimation of enhancing protein stability due to the presence of long-range links between residues. As a stability measure (which defines the hardness against the thermal influence), one could use the temperature T_c of the phase transition from the state with the non-zero order parameter into the state where it equals

Fig. 2 Temperature dependencies of the order parameter $\varepsilon(T)$ for the protein, consisting in the core and shell with volume fractions v_1, v_2 and characterized by the average numbers n_{01}, n_{02} of links between residues and their energies J_1 and J_2 in the core and shell, respectively. We have assumed: $v_1 = 0.7, v_2 = 0.3, n_{01} = 8.5, n_{02} = 5, \sigma = 2$. Calculations have been performed for two cases: $J_2/J_1 = 1$ and $J_2/J_1 = 1.5$. *Insert*; temperature dependencies of the derivative $d\varepsilon/dT$, reproducing the external view of experimental temperature dependencies of the heat capacity

zero. We will show that with increasing the long links' concentration p the temperature T_C raises as well.

The physical reason of this increase could be readily understood with a simple example of the one-dimensional Ising magnetic system. For the regular linear chain, the ordered state is impossible and $T_c = 0$. In fact, the partition of the spin chain into two domains results in increasing the system energy by $2J$. However, any of N links could serve as a boundary between domains that corresponds to increasing entropy by $k \ln N$. Thus, the variation of the free energy, being equal to $\Delta F = 2J - kT \ln N$, is always negative for sufficiently large N . This means that the partition into domains is energetically favorable. However, if there are long-range links, the possibility to choose the position of the domain boundary is strongly limited—this boundary could not be positioned between ends of the long-range link (the simple rule: domain boundary and long link cannot be crossed). The number of links, suitable for such a boundary, diminishes by approximately $pN \gg 1$ times, as compared with the original links' number N and equals $\sim N/pN = 1/p$. Now the free energy variation is $\Delta F \sim 2J/kT - \ln(1/p)$. It is positive (i.e., partition into domains is unfavorable) if $T < T_c^{(1D)}$, where

$$kT_c^{(1D)} \sim 2J / |\ln p|. \quad (15)$$

Hence, long-range links promote ordering in the one-dimensional system. This conclusion coincides with results corresponding to the exact solution for some specific cases of one-dimensional small-world networks [33].

Analogous arguments, going back to the paper [34], allow to find dependencies $T_C(p)$ for two- and three-dimensional small-world networks. In the first case, generating domains with the boundary of length ℓ (in units equal to the distance between nearest neighbors) leads to the formation of ℓ incorrectly mutually oriented pairs of residues that increases the system energy by $2\ell J$. To calculate the entropy of the system with such domains, one needs to estimate the number G of ways to trace the boundary of length ℓ . In each node, such a boundary could choose one out of three directions, $G \sim 3^\ell$. (The boundary should be closed, so that number is slightly overestimated. However, for large ℓ inaccuracy is

Fig. 3 Data array revealing the correlation between the protein melting temperature T_m and the fraction p of disulphide bonds. The shaded area of data positioning is a helpful guide. The solid line is the result of linear regression analysis and the dashed line is the theoretical dependence (17)

negligible [35].) Thus, generating the domain boundaries results in variation of the free energy equal to $\Delta F \approx 2\ell J - kT \ln G = \ell(2J - kT \ln 3)$. Original (single-domain) state is favorable ($\Delta F > 0$) if $T < T_c^{(2D)}$, where $kT_c^{(2D)} = 2J/\ln 3 = 1.82J$. Long-range links diminish the number of possible boundaries: they cannot travel between nodes connected by such a link, since that would increase the system energy. For this reason, there appear nodes where the boundary could choose for its continuation not three but only two directions. The number of such nodes is on the order of $(p\ell^2)^{1/2} = \ell\sqrt{p}$, so that number of possible the boundaries of length ℓ lowers down to $G_p \sim 3^{\ell-\ell\sqrt{p}} \cdot 2^{\ell\sqrt{p}}$. The corresponding free-energy variation equals $\Delta F \approx 2\ell J - kT \ln G_p = \ell \{2J - kT [\ln 3 - p^{1/2} \ln(3/2)]\}$. It is positive at $T < T_c^{(2D)}$, where

$$kT_c^{(2D)} = \frac{2J/\ln 3}{1 - [1 - (\ln 2/\ln 3)]\sqrt{p}} \rightarrow 1.82J(1 + 0.37\sqrt{p}) \quad (p < 1). \quad (16)$$

Hence, with increasing p the phase transition temperature grows.

A similar approach could also be used in the three-dimensional case where one needs to consider the number of possible three-dimensional domains with a fixed surface area. The phase-transition temperature turns out to be equal to [36]

$$kT_c^{(3D)} = \frac{4.5J}{1 - 0.29p^{2/3}} \rightarrow 4.5J(1 + 0.29p^{2/3}) \quad (at \ p < 1). \quad (17)$$

Thus, the presence of long-range links in the protein globule promotes its stability whose measure is the phase transition temperature (denaturation temperature). When increasing the percent p of such links, the temperature T_c grows, so that in the three-dimensional case it increases from $4.5J$ (at $p = 0$) up to $4.8J$ at $p = 0.1$. For $J \sim 0.2kT$ this leads to $\Delta T_c^{(3D)} \sim 20$ K, which evidences a significant enhancement of the thermal stability of the protein globule due to the mechanism considered. Notice, however, that it becomes effective when the probable number of these long-range links $pN \gg 1$, that is at $p \gg 1/N \sim 0.03$ (at $N \sim 100$).

To examine how the conclusion corresponds to reality, we used the known fact [37, 38] that inter-protein disulfide bonds enhance protein thermostability. In the context of our model, disulfide bonds are long-range links, providing the high temperature of the phase transition in the protein small-world network, which we associate with the melting point T_m . In Table 1 (see Appendix) are listed some proteins with their melting temperatures T_m along

with the total number N_R of residues, the number N_S of disulfide bonds and the effective fraction $p = N_S/N_R$ of those bonds. The corresponding data array (p, T_m) is represented in Fig. 3. Though the number of disulfide bonds is not a single factor influencing the proteins' thermostability, the presented diagram allows not only to highlight the qualitative tendency but also to give a quantitative estimate of the $T_m(p)$ dependence. The standard recipe of this estimation is linear regression data analysis resulting in the solid line in Fig. 3.

4 Conclusions

The above introduced order parameter ε is not exactly specified: it is associated with some qualitative idea about deviating a protein's three-dimensional structure from its equilibrium state. The skeleton of the protein macromolecule shakes continuously and the amplitude of these fluctuations increases with the temperature rise until the fraction of not quite properly positioned amino acid residues amounts to half of their total number (at $\varepsilon = 0$). This is just the phase transition corresponding to the globule melting. The considered model describes the temperature dependence of the order parameter, which is not clearly defined. Intuitively, we associate it, on the one hand, with the extent of distorting the protein macromolecule structure and on the other hand with the energy increment that arises at such a distortion. The temperature derivative $d\varepsilon/dT$ of the order parameter could serve as a measure of the system's heat capacity measured by the calorimetric method.

From the generalized mean-field model, it follows that the larger the average number of weak links per one amino acid residue in the globule, the higher the temperature of the phase transition. However, this temperature depends weakly on the width of the distribution function of those links' numbers. This temperature is determined by the relation $kT_c \approx (n_0 - 1)J$ (see Fig. 1), wherefrom it follows that the energy variation due to moving the residue from the right position to the wrong one equals $J \sim kT_c/n_0 \sim 0.1$ cal/mol, which is much less than the thermal energy. This provides the liability and the internal flexibility of the protein macromolecule that are necessary for its functionality [29].

Residues positioned near the globule surface have, on average, a lower number of links than those situated in the core. Therefore, with increasing temperature, the globule shell is melted first, and then the core melts [31]. The difference of the relevant melting temperatures equals $\Delta T_c \sim (n_{01}J_1 - n_{02}J_2)/k \sim 30$ K (indexes 1 and 2 refer to the globule core and shell, respectively). Since the shell amounts to 20–30% of the protein volume, both transitions could be easily observed by the calorimetry method.

The temperature stability of the protein globule is defined by its skeleton being, from a mathematical point of view, the graph whose nodes coincide with centers of amino acid residues and edges correspond to weak links between those that are situated near each other in the globular state. Protein graphs (networks), generated in such a way, are neither regular nor random ones. They have some intermediate place between these two limiting network types and are referred to as so-called small-world networks. In such networks, there is a noticeable fraction (~ 3 –5%) of long-range links. They randomly connect residues that are distant from each other in the original protein chain but become neighbors in the protein folded in the globule. Just these links in proteins give additional stability (as compared with the regular systems) against the denaturing external factors. For example, the stabilizing effect of long-range links leads to increasing globule melting temperature by ~ 10 –30 K. Disulfide bonds in thermophilic proteins could play the role of these links [22].

Acknowledgement This work has been supported by Grants ## 12-02-00550, 11-02-00363a of the Russian Foundation of Basic Researches.

Appendix

Table 1 Melting temperatures T_m of some proteins with the total number N_R of their residues, the number N_S of disulfide bonds and the effective fraction $p = N_S/N_R$ of those bonds*

Protein	T_m , K	N_R	N_S	$p = N_S/N_R$
α -amilase	335	240	T0	0
α -chymotrypsin	328	240	T2	0.008
β -trypsin	365	223	6	0.027
ABP	326	201	1	0.005
Albumin	366	609	17	0.028
Apo- α -lactalbumin	345	142	4	0.028
Barnase	336	110	0	0
BPTI	369	58	3	0.05
Bovine hemoglobin	331	145	0	0
Bovine serum albumin	358	607	17	0.028
Carbonic anhydrase	342	130	0	0
CI-2	343	130	0	0
Cytochrome <i>c</i> , horse	354	105	0	0
Eglin	357	70	0	0
Ferredoxin	398	98	3	0.031
Insulin, human	357	110	3	0.027
Interleukin-1 β	335	72	2	0.028
Lysozyme, chicken	365	129	4	0.031
Myoglobin, carp	341	147	0	0
Myoglobin, horse	355	154	0	0
Ovalbumin	364	386	1–2	0.004
Ovomucoid	372	186	9	0.048
Papain	360	212	4	0.019
Parvalbumin, carp	366	108	0	0
Pepsinogen	352	384	3	0.008
PGK	373	449	6	0.013
Plasminogen, fr K4	338	840	12	0.014
Protein G	351	524	12	0.013
Ribonuclease A	337	124	4	0.032
RNase	370	147	3	0.02
Rubredoxin	392	174	4	0.023
Staph nuclease	337	276	1	0.036
Tendamistat	351	74	2	0.027

*Information from the Protein Data Bank UniProt (<http://www.uniprot.org/>) has been used

References

1. Barabási, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004)
2. Newman, M.E.J.: The structure and function of complex networks. *SIAM Rev.* **5**, 167–256 (2003)
3. Atılgan, C., Okan, O.B., Atılgan, A.R.: Network-based models as tools hinting at non-evident protein functionality. *Ann. Rev. Biophys.* **41**, 205–225 (2012)
4. Miyazawa, S., Jernigan, R.L.: Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644 (1996)
5. Bagler, G., Sinha, S.: Assortative mixing in protein contact networks and protein folding kinetics. *Bioinformatics* **23**, 1760–1767 (2007)

6. Atilgan, A.R., Akan, P., Baysal, C.: Small-world communication of residues and significance for protein dynamics. *Biophys. J.* **86**, 85–91 (2004)
7. Vendruscolo, M., Dokholyan, N.V., Paci, E., Karplus, M.: Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev.* **65**, 061910 (2002)
8. Gaci, O., Balev, S.: Motif prediction in amino acid interaction networks, vol. I. In: *Proc. World Congress on Eng. and Comp. Sci.*, San Francisco, USA (2009)
9. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. *Nature (London)* **393**, 440–442 (1998)
10. Milgram, S.: The small-world problem. *Psychol. Today* **1**, 61–67 (1967)
11. Jiao, X., Chang, S.: Scoring function based on weighted residue network. *Int. J. Mol. Sci.* **12**, 8773–8786 (2011)
12. Plaxco, K.W., Simons, K.T., Baker, D.: Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998)
13. Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D., Finkelstein, A.V.: Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12**, 2057–2062 (2003)
14. Finkelstein, A.V., Ivankov, D.N., Garbuzynskiy, S.O., Galzitskaya, O.V.: Understanding the folding rates and folding nuclei of globular proteins. *Curr. Protein Pept. Sci.* **8**, 521–536 (2007)
15. Herrero, C.P.: Ising model in small-world networks. *Phys. Rev. E* **65**, 66110 (2002)
16. Grosberg, A.Y., Khokhlov, A.R.: *Statistical Physics of Macromolecules*. Springer (1994)
17. Landau, L.D., Lifshitz, E.M.: *Statistical Physics*. Butterworth-Heinemann (1980)
18. Kittel, C.: *Introduction to Solid State Physics*. Wiley (2004)
19. Chandrasekhar, S.: *Liquid Crystals*. Cambridge Univ. Press (1994)
20. Birstein, T.M., Ptitsy, O.B.: *Conformation of Macromolecules*. Nauka, Moscow (in Russian) (1964)
21. Mallick, P., Boutz, D.R., Eisenberg, D., Yeates, T.O.: Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proc. Natl. Acad. Sci.* **99**, 96799684 (2002)
22. Jorda, J., Yeates, T.O.: Widespread disulfide bonding in proteins from thermophilic Archaea. *Hindawi Publ. Corp. Archaea*, vol. 2011, 9 p. Article ID 409156. doi:[10.1155/2011/409156](https://doi.org/10.1155/2011/409156)
23. Meilikhov, E.Z., Farzetdinova, R.M.: Generalized mean-field theory for Ising spins in small-world networks. *Phys. Rev. E* **71**, 046111 (2005)
24. Berman, H., Henrick, K., Nakamura, H.: Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003)
25. Huang, J., Kawashima, S.I., Kanehisa, M.: New amino acid indices based on residue network topology. *Gen. Inf.* **18**, 152–161 (2007)
26. Romero, P., Obradovic, Z., Kissinger, C., Villafrance, J.E., Dunker, A.K.: Identifying disordered regions in proteins from amino acid sequence. In: *Proceedings of the International Conference on Neural Networks*, pp. 91–95 (1997)
27. Scopes, R.K.: *Protein Purification: Principles and Practice*, 3rd edn., pp. 98–100. Springer (1994)
28. Privalov, P.L., Dragan, A.I.: Microcalorimetry of biological macromolecules. *Biophys. Chem.* **126**, 16–24 (2007)
29. Volkenshtein, M.V.: *Biophysics*. Mir, Moscow (1983)
30. *Nanoscale Materials in Chemistry*, 2nd edn. Klabunde, K.J., Richards, R.M. (eds.). Wiley (2009)
31. Zhou, Y., Vitkup, D., Karplus, M.: Native proteins are surface-molten solids: application of the Lindemann criterion for the solid versus liquid state. *J. Mol. Biol.* **285**, 1371–1375 (1999)
32. Privalov, P.L.: Stability of proteins. Proteins which do not present a single cooperative system. *Adv. Protein Chem.* **35**, 1–104 (1982)
33. Lopes, J.V., Pogorelov, Y.G., Lopes dos Santos, J.B.M., Toral, R.: Exact solution of Ising model on a small-world network. *Phys. Rev. E* **70**, 026112 (2004)
34. Peierls, R.: On Ising model of ferro magnetis. *Proc. Camb. Philol. Soc.* **32**, 477–481 (1936)
35. Griffiths, R.B.: Peierls proof of spontaneous magnetization in a two-dimensional Ising ferromagnet. *Phys. Rev.* **136**, A437–A439 (1964)
36. Meilikhov, E.Z., Farzetdinova, R.M.: Curie temperature for small-world Ising systems of different dimensions. *J. Magn. Magn. Mater.* **300**, 254–256 (2006)
37. Birgelson, J.D., Wolynes, P.G.: Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 7524–7528 (1987)
38. Birgelson, J.D., Wolynes, P.G.: Intermediates and barrier crossing in random energy model with application to protein folding. *J. Phys. Chem.* **93**, 6902–6915 (1989)