Favorable domain size in proteins

Dong Xu[1]* and Ruth Nussinov[1,2]

Background: It has been observed that single-domain proteins and domains in multidomain proteins favor a chain length in the range 100–150 amino acids. To understand the origin of the favored size, we construct an empirical function for the free energy of unfolding versus the chain length. The parameters in the function are derived by fitting to the energy of hydration, entropy and enthalpy of unfolding of nine proteins. Our energy function cannot be used to calculate the energetics accurately for individual proteins because the energetics also depend on other factors, such as the composition and the conformation of the protein. Nevertheless, the energy function statistically characterizes the general relationship between the free energy of unfolding and the size of the protein.

Results: The predicted optimal number of residues, which corresponds to the maximum free energy of unfolding, is 100. This is in agreement with a statistical analysis of protein domains derived from their experimental structures. When a chain is too short, our energy function indicates that the change in enthalpy of internal interactions is not favorable enough for folding because of the limited number of inter-residue contacts. A long chain is also unfavorable for a single domain because the cost of configurational entropy increases quadratically as a function of the chain length, whereas the favorable change in enthalpy of internal interactions increases linearly.

Conclusions: Our study shows that the energetic balance is the dominant factor governing protein sizes and it forces a large protein to break into several domains during folding.

Addresses: [1]Laboratory of Experimental and Computational Biology, IRSP, SAIC Frederick, NCI-FCRDC, Frederick, MD 21702-1201, USA. [2]Sackler Institute of Molecular Medicine, Tel Aviv University, Tel Aviv 69978, Israel.

*Present address: Computational Biosciences, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6480, USA.

Correspondence: Ruth Nussinov
E-mail: ruthn@ncifcrf.gov

Key words: domain, entropy, free energy, protein folding, statistical analysis

Received: **15 September 1997**
Revisions requested: **08 October 1997**
Revisions received: **15 October 1997**
Accepted: **22 October 1997**

Published: **27 November 1997**
http://biomednet.com/elecref/1359027800300011

Folding & Design 27 November 1997, **3**:11–17

Introduction

It has been widely recognized that proteins are often divided into domains having 100–150 amino acids [1–4]. Given the regular size of these protein building blocks, a periodicity in protein sizes is expected. Indeed, clustering of molecular sizes around multiples of a unit size, ~14 kDa, has been observed in *Escherichia coli* proteins by gel electrophoresis [5]. A statistical analysis of a large collection of nonredundant protein sequences indicates that proteins consist of sequence units with characteristic lengths of 125 residues for eukaryotes and 150 residues for prokaryotes [6]. It has been proposed that multidomain proteins may have evolved from proteins having single domains via domain insertion [7]. For example, mammalian aspartic proteases, which contain two structurally homologous lobes, are suggested to have arisen during evolution from a homodimer enzyme via gene duplication and fusion [8,9]. Similarly, the reverse transcriptase may have evolved from domain fusion of ancestors of the type I ribonuclease H and the polymerase domain [10]. During the folding process, large proteins are thought to form several stable collapsed hydrophobic folding units or domains, which then assemble [11,12]. The folding of each stable domain in a multidomain protein is similar to the folding of a single-domain protein [13].

The rather uniform size of this structural unit suggests that a general principle such as geometrical or physical optimization at the DNA or protein level is responsible [5]. Berman *et al.* [6] proposed a possible recombinational origin of the domain structure. Because the optimal number of residues in protein domains corresponds to the optimal size for circularization of DNA circles [14], the authors suggest that a DNA circle could be an elementary recombinational unit in the early evolution of protein-coding sequences. Although this hypothesis may be valid, the optimization at the gene level during evolution has been under the pressure of the genes' products (i.e. proteins). Hence, it is reasonable to assume that there is a fundamental base in energetics for protein domains to favor a particular size. Dill developed a lattice model to account for the energetics of protein folding [15]. The model predicts an optimal chain length for maximal protein stability. If a chain is too short, the protein would have too little interior volume to form enough favorable contacts between hydrophobic residues. If a chain is too long, the protein is forced to bury many unfavorable hydrophilic residues in the interior of the protein, given a particular ratio between the number of hydrophobic and hydrophilic residues. But because the theory indicates that the optimal domain size can increase rapidly with increasing the ratio between hydrophobic and hydrophilic

residues, it failed to explain why during evolution favorable large protein domains have not been generated by increasing the fraction of hydrophobic residues in a chain.

In order to study the folding thermodynamics as a function of protein chain length it is important and feasible to employ a more realistic model of protein energetics than the highly simplified lattice model. Such a study will provide a quantitative rationale for protein domains to have a preferable size. It will also shed some light on domain formation in large proteins. Here, we provide a statistical description of the energy terms as a function of chain length, matched through thermodynamic data characterized by Makhatadze and Privalov [16]. It is not our intention to derive a universal free energy function for the chain length for individual proteins because folding thermodynamics not only depends on the chain length, but also on other factors such as the shape of the protein and its disulfide bonds. Nevertheless, the statistical energy function of chain length reflects the average trend in protein energetics. In particular, as we show below, the optimal chain length in protein domains, predicted from the statistical energy function, is in agreement with the observation on experimental protein structures.

In this paper, we first introduce the theory and the methods. Then we present the results of the derived energy function and its prediction of optimal domain size, compared with the statistical analysis of protein structures. Finally we discuss the validity and implications of our results.

Results

Theory and methodology

We first describe our energy terms as a function of chain length and the assumptions that they entail. Next we provide the data used in fitting the energy functions. Finally we introduce the method employed in the statistical analysis of the protein structures.

Thermodynamics of protein unfolding

Let us consider a protein unfolding at temperature T. The overall change in free energy for the protein unfolding, ΔG, can be described by [16]:

$$\Delta G = \Delta G_{hyd} + \Delta H_{int} - T\Delta S_{conf} \quad (1)$$

where ΔG_{hyd} is the change in the free energy of hydration on unfolding, ΔH_{int} is the change in the enthalpy of internal interactions on unfolding in the gaseous phase, and ΔS_{conf} is the configurational contribution to the entropy change during the unfolding. The sign of ΔG governs the folding process. If $\Delta G > 0$, the peptide chain can fold into a folded protein; if $\Delta G < 0$, the chain cannot fold.

ΔG_{hyd} is the free energy change of unfolding due to the transfer of the protein from the gaseous phase to water. If the change in free energy is $\Delta G'$ in the gaseous phase for the hypothetical unfolding process that forms the same native protein structure then:

$$\Delta G_{hyd} = \Delta G - \Delta G' \quad (2)$$

where ΔG is the change in free energy for the native protein unfolding, as used in Equation 1. The value of ΔG_{hyd} is proportional to ΔASA, the change in the solvent accessible surface area on unfolding [16]. For globular proteins, ΔASA, in units of Å^2, can be written as [17]:

$$\Delta ASA = 1.48\, M_w - 6.3\, M_w^{0.73} \quad (3)$$

where M_w is the molecular weight of the protein in units of Dalton. The average molecular weight of the residues in globular proteins is 119.40 Da [18]. Hence, M_w can be approximated by:

$$M_w = 119.40n \quad (4)$$

where n is the number of residues. Taken together:

$$\Delta G_{hyd} = a\,(176.71n - 206.80n^{0.73}) \quad (5)$$

where a is a constant to be determined.

The change in the enthalpy of intramolecular interactions on unfolding in the gaseous phase, ΔH_{int}, can be written as:

$$\Delta H_{int} = b\,n + c \quad (6)$$

where b and c are constants.

As shown in [16], the specific configurational entropy of protein unfolding, S_{conf}/n increases with increasing n. Thus:

$$S_{conf}/n = f\,n + g \quad (7)$$

where f and g are constants. We shall come back to justify the functional forms of ΔH_{int} and S_{conf}/n.

Thermodynamic data

To determine the parameters a, b, c, f and g in Equations 5–7, we have employed a data set of nine single-domain proteins and we use ΔG_{hyd}, ΔH_{int} and ΔS_{conf}, all of which were derived from the experimental free energy of unfolding ΔG [16]. The physiological temperature (37°C) is used:

$$T = 310K \quad (8)$$

The thermodynamic quantities in [16] were measured at 25°C and 50°C. Because the interval between 25°C and 50°C is small, we linearly extrapolated a quantity at 37°C [$Q(37)$] from the corresponding values at 25°C [$Q(25)$] and 50°C [$Q(50)$]:

$$Q(37) = \frac{13}{25}\,Q(25) + \frac{12}{25}\,Q(50) \quad (9)$$

The values of ΔG_{hyd}, ΔH_{int} and $\Delta S_{conf}/n$ at 37°C in nine proteins are listed in Table 1.

Domain size

Islam *et al.* [19] assigned the domains in the 284 non-redundant protein chains based on inter-residue contacts

Table 1

Thermodynamics characteristics of the studied proteins at 37°C.

Protein	n	ΔG_{hyd} (kJ mol^{-1})		ΔH_{int} (kJ mol^{-1})		$\Delta S_{conf}/n$ (J K^{-1} mol^{-1})	
BPTI	58	−1784	(−1467)	2349	(2345)	46.92	(49.09)
Ubiquitin	76	−1776	(−2009)	2960	(3246)	48.69	(50.49)
RNase T1	104	−2652	(−2877)	4338	(4648)	51.57	(52.68)
Cytochrome *c*	104	−3408	(−2877)	5282	(4648)	57.14	(52.68)
Barnase	110	−2990	(−3065)	4939	(4949)	56.14	(53.15)
RNase A	124	−3544	(−3509)	5658	(5650)	54.66	(54.24)
Lysozyme	129	−3824	(−3669)	6151	(5900)	56.27	(54.63)
Interleukin 1β	153	−3835	(−4442)	6296	(7102)	51.52	(56.51)
Myoglobin	153	−4875	(−4442)	7618	(7102)	57.06	(56.51)

n, Number of amino acids; ΔG_{hyd}, change in free energy of hydration on unfolding; ΔH_{int}, change in the enthalpy for internal interactions on unfolding; ΔS_{conf}, configurational contribution to the entropy for protein unfolding. The numbers that are not in brackets derive from the data in [16] at 37°C; the numbers in brackets were calculated using Equations 5–7.

in experimental protein structures. There are 197 single-domain proteins and 202 domains in multidomain proteins. We have employed the sizes of the domains in their domain assignment. Assume there are M(i) domains, where i is the number of residues and i = 1, 2, 3, ..., N. The distribution of protein domains around a domain of m residues in length, P(m), can be calculated by:

$$P(m) = \frac{1}{(2v+1)K}\left[\sum_{i=m-v}^{i=m+v} M(i)\right] \quad (10)$$

where K is the total number of protein domains, and $2v + 1$ is the window size, in number of amino acids.

We now present the empirical free energy of unfolding as a function of the number of residues n. From this function, we will derive the optimal length of amino acids in a protein domain and compare it with the statistical analysis of experimental protein structures.

Matching the parameters

The data in Table 1 are matched by Equations 5–7 through the least squares fit. The parameters obtained are as follows:

$$a = -0.2353 \text{ kJ mol}^{-1} \quad (11)$$

$$b = 50.07 \text{ kJ mol}^{-1} \qquad c = -559.2 \text{ kJ mol}^{-1} \quad (12)$$

$$f = 0.07809 \text{ J K}^{-1} \text{ mol}^{-1} \qquad g = 44.56 \text{ J K}^{-1} \text{ mol}^{-1} \quad (13)$$

Figure 1 and Table 1 show the comparison between the experimental values and the fitting results derived from Equations 5–7 and the parameters given in Equations 11–13. The correlation coefficients between the values in [16] and those calculated by our energy function are 0.934 for ΔG_{hyd} and 0.964 for ΔH_{int}, both values indicating a very good fit. The correlation coefficient of $\Delta S_{conf}/n$ is 0.655, which is not as good as the values for ΔG_{hyd} and ΔH_{int} but is still significant. The relatively weak correlation of $\Delta S_{conf}/n$ is due to the less sensitive dependence of $\Delta S_{conf}/n$ on n. Overall, the quality of the fittings shows the validity of the functional forms in Equations 5–7. In particular, there is a constant term c in ΔH_{int} as shown in Equation 6. The functional form of ΔH_{int} can be understood from a heuristic illustration. ΔH_{int} describes the change in enthalpy of internal interactions on unfolding in the gaseous phase. If a peptide chain is too short, it cannot form enough inter-residue contacts for intramolecular interactions to contribute a positive ΔH_{int}. This is reflected in the negative constant c in ΔH_{int}, which ensures the minimum number of residues to establish stable interactions (i.e. $\Delta H_{int} > 0$ only if $n > 11$; $\Delta H_{int} < 0$ in the range of $n \leq 11$ is an artifact of the fitting, but it is irrelevant to our subject). Once the stable interactions are established, the contribution of a particular residue to ΔH_{int} is mostly from its neighboring residues. Because the density is homogenous in proteins, the number of neighboring residues is basically constant. Hence, it is not surprising that ΔH_{int} linearly correlates with the number of residues n.

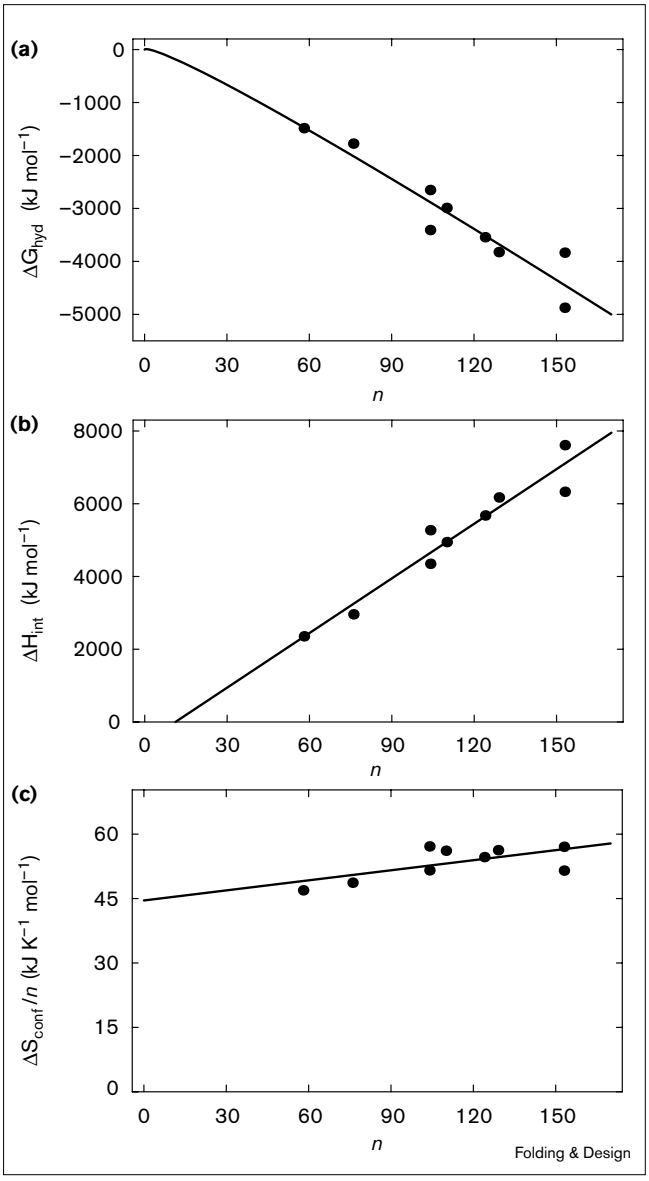
Optimal domain size derived from the fitting function

By using Equation 1 and Equations 5–7 together with the fitting parameters, the free energy of unfolding can be written as:

$$\Delta G(n) = -0.02422n^2 - 5.333n + 48.60n^{0.73} - 559.2 \text{ (kJ mol}^{-1}\text{)} \quad (14)$$

Figure 2a shows $\Delta G(n)$ and its three energy components, ΔG_{hyd}, ΔH_{int} and $-T\Delta S_{conf}$, derived from Equations 5–7. ΔG_{hyd} and $-T\Delta S_{conf}$ destabilize the folded protein. ΔH_{int} contributes to the protein stability when $n > 11$. The three components are very large compared to $\Delta G(n)$ itself.

Figure 1



The data derived from [16] (filled circles) and theoretical fitting curves of Equations 5–7 (lines) as a function of the number of residues (n), at 37°C. **(a)** The change in the free energy of hydration on unfolding, ΔG_{hyd}; **(b)** the change in the enthalpy of internal interactions on unfolding, ΔH_{int}; and **(c)** the configurational component of the change in entropy of protein unfolding per residue ($\Delta S_{conf}/n$).

Figure 2b indicates that there is only a narrow region of n (i.e. $n_{min} < n < n_{max}$) in which $\Delta G(n) > 0$. $\Delta G(n)$ reaches its maximum value at $n = n_{opt}$. The values obtained for n_{min}, n_{max} and n_{opt} are:

$$n_{min} = 60.4, \quad n_{max} = 143.1, \quad n_{opt} = 100.4 \qquad (15)$$

Hence, the optimal length of a single-domain protein is 100 residues, and folded single-domain proteins are stable in the range 61–143 residues.

Statistical analysis of protein domains

The distributions of protein domain sizes are shown in Figure 3. Figure 3a compares the distributions of single-domain protein chains with those of domains in multi-domain chains, over intervals of 50 amino acids. The figure illustrates that the two distributions are similar. When the two distributions are combined, we obtain the average domain size:

$$n_{ave} = 148 \qquad (16)$$

To determine the optimal n value (n_{dom}) for the maximum $P(n)$, we have calculated $P(n)$ for contiguous n with a window size of 21, as shown in Figure 3b. The optimal n value is:

$$n_{dom} = 106 \qquad (17)$$

Figure 3b indicates that $P(n)$ is asymmetrical.

Discussion

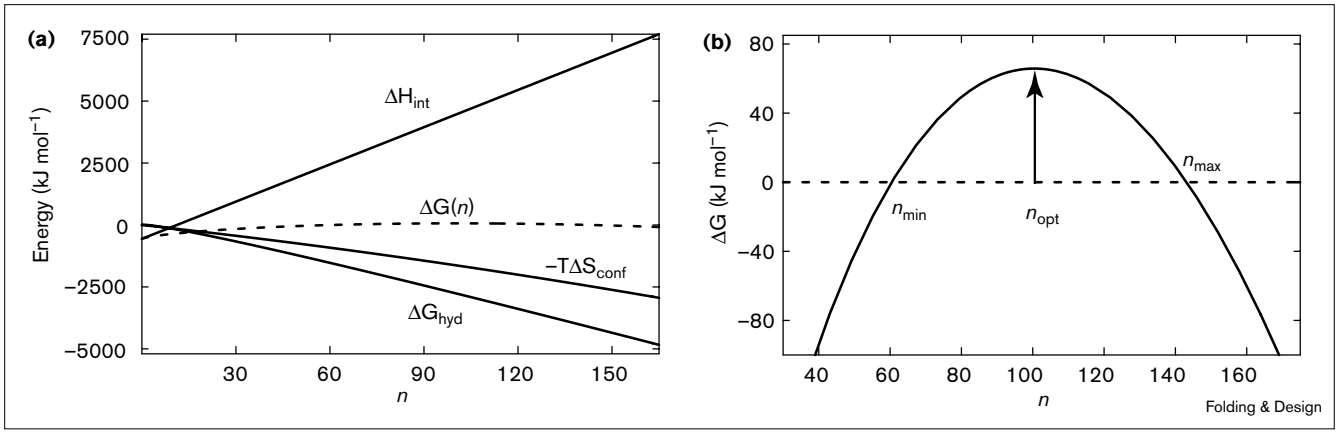
We now discuss the validity of our energy function, the optimal chain length of proteins based on the free energy function, and the entropic contribution, which is crucial for determining the range of favorable chain lengths in a domain. We also analyze the thermodynamic origin of the division of a large protein into several domains. Finally, we address some limitations of our study and possible further explorations along our approach.

Energy functions

We have derived the energy functions for the change in free energy of hydration for unfolding (ΔG_{hyd}), the change in enthalpy of internal interactions for unfolding (ΔH_{int}), and the configurational contribution to the entropy of protein unfolding (ΔS_{conf}) at 37°C for nine single-domain proteins. There are alternative fitting approaches establishing the relationship between the energetics of the folding of a protein sequence and its corresponding protein structure. For example, one may fit the energy terms by the change of the hydrophobic and hydrophilic surface areas during unfolding [20] and then convert the surface areas to the number of residues n. Such indirect conversions may, however, lead to an unreliable dependence of n due to the strong correlation between the hydrophobic surface area and the hydrophilic one. For studying n-dependent energetics, it is important to fit the experimental thermodynamic data directly to n, as we have done here. Our energy functions do not presume a dielectric constant of the protein, a particular shape of its domains, or a particular ratio between the number of hydrophobic and hydrophilic residues.

Our assumed energy functions fit well with the data of ΔG_{hyd}, ΔH_{int} and $\Delta S_{conf}/n$ in nine proteins. On the other hand, the absolute value of the free energy of unfolding is much less than its components ΔG_{hyd}, ΔH_{int} and $-T\Delta S_{conf}$. Typical values for the free energy of unfolding are in the order of tens of kJ mol^{-1}, whereas the three components

Figure 2



Energy terms as a function of the number of residues n, at 37°C. **(a)** $\Delta G(n)$ based on Equation 14 (dashed line) and its components (ΔG_{hyd}, ΔH_{int} and $-T\Delta S_{conf}$; solid lines) based on Equations 5–7. **(b)** The solid line is a magnification of (a) for $\Delta G(n)$. The dashed line marks $\Delta G = 0$. n_{min} and n_{max} are the cross points between the two lines (i.e. $\Delta G(n) > 0$ for $n_{min} < n < n_{max}$) and n_{opt} is the optimal value at which $\Delta G(n)$ reaches its maximum.

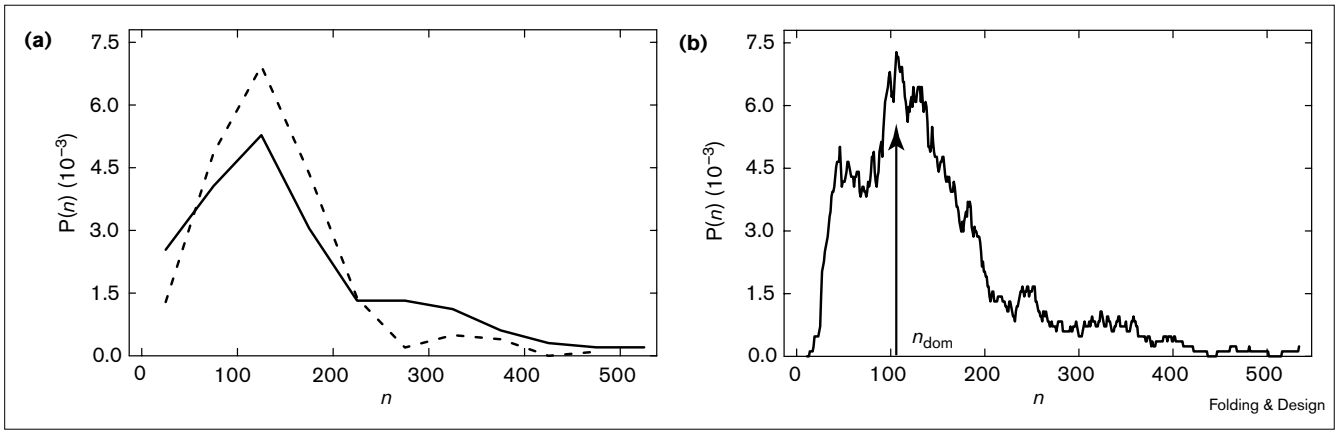
are usually two orders higher. Hence, one cannot use our free energy function to calculate the free energy of unfolding for an individual protein. Nevertheless, the free energy function, derived from the three components, reflects well the statistical profile of domain stability versus chain length, as well as the involvement of each free energy component.

Optimal chain length in proteins

Our free energy function predicts that the optimal number of residues in a single-domain protein is 100. This is in agreement with a statistical analysis of protein domains, which indicates an optimal length of 106. The quantitative agreement substantiates the notion that the 'foldability' of a chain governs the preferred length of protein domains.

The optimal free energy of unfolding at n = 100 originates from the balance of three components, ΔG_{hyd}, ΔH_{int} and $-T\Delta S_{conf}$. Both ΔG_{hyd} and $-T\Delta S_{conf}$ destabilize folded proteins. As n becomes larger, ΔH_{int} gradually overcomes the two unfavorable free energy components and drives the total free energy to be favorable for folding. If a peptide chain is too short, there are insufficient inter-residue contacts to have a large enough ΔH_{int} for a stable folded protein, similar to the scenario depicted by Dill [15]. When

Figure 3



Domain length distribution P(n) as a function of the number of residues n. **(a)** P(n) for residues 1–50, 51–100, 101–150, ... with a window of 50 amino acids. The solid line represents the single-domain chains (197 chains) and the dashed line represents the domains in the multidomain chains (202 domains). **(b)** P(n) versus the domain size of n residues, for the domains in both the single-domain chains and the multidomain chains (399 domains). P(n) is calculated for a window size of 21 (i.e. in the range of $n \pm 10$). n_{dom} indicates the n value at which P(n) reaches its maximum.

$n > 100$, because $-T\Delta S_{conf}$ has a quadratic form as a function of n, $-T\Delta S_{conf}$ decreases the free energy of unfolding faster with respect to n than ΔH_{int} increases the free energy of unfolding; ΔH_{int} increases only linearly versus n. Hence, a single-domain protein cannot be too large. This explanation differs from Dill's suggestion that a long chain is unfavorable for a single-domain protein because it has to bury many unfavorable hydrophilic residues in the interior [15]. Compared with Dill's theory, our model is based on a more realistic energy function rather than a simplified lattice model, and our quantitative prediction does not require any presumed composition between the hydrophobic residues and the hydrophilic residues.

The dominant factor in governing the size of single-domain proteins is the quadratic form of $\Delta S_{conf}(n)$. If $\Delta S_{conf}(n)$ were linear, the larger (or the smaller) the protein size the better. The origin of the n^2 dependence of $\Delta S_{conf}(n)$ has been suggested to account for the surface residues, which are less compact than the internal residues [16]. The number of surface residues is proportional to $n^{2/3}$ rather than n^2, however. We therefore propose another mechanism for the n^2 dependence in $\Delta S_{conf}(n)$. As a protein folds, the first entropy loss is the restraint in the movement of each residue, for both its backbone and its sidechain. This term is proportional to n. In addition, a sequence has to fold into a unique configuration. There is only a limited number of energetically allowed conformational categories or protein folds for a single domain [21,22]. The protein folds impose restraints on the inter-residue distances. The longer a sequence, the more restraints it has. This additional cost in configurational entropy during folding increases dramatically as n increases, and may result in the n^2 dependence in $\Delta S_{conf}(n)$.

Our free energy function predicts that folded single-domain proteins only stabilize in the range 61–143 residues. Although most protein domains in the statistical analysis of protein structures fall into this range, there are some single-domain proteins beyond it. We suggest the following reasons for the discrepancy. First, there may be errors originating from the data in [16] and the fitting of the energy terms, so that the free energy function is not accurate enough. Second, some of the proteins in our statistical analysis are transmembrane or membrane-bound proteins. For example, the M segment of the transmembrane protein photosynthetic reaction center has 333 residues; and cytochrome P450, which is a membrane-bound protein, has 457 residues. Because our energy function was derived from globular proteins, it cannot describe transmembrane or membrane-bound proteins. Third, disulfide bonds can reduce the entropy cost so that the range of favorable chain length can be widened. Small proteins, in particular, usually have disulfide bonds ([18]; e.g. there are three disulfide bonds in the protein BPTI with 58 residues), but very large proteins also have disulfide bonds (e.g. there are nine disulfide bonds in the protein neuraminidase with 388 residues). Fourth, some large single-domain proteins consist of loops which do not pack compactly around the protein core. For example, in the large protein purine nucleoside phosphorylase (289 residues; PDB code 1ula), the loop regions 59–65, 248–256 and 284–289 hang around the protein core and only loosely connect to other parts of the protein. Such uncompact packing can also reduce the entropic cost of protein folding so as to increase the upper bound of single-domain protein size. Finally, the method of Islam *et al.* [19] may not be sensitive enough to find all the possible domains in proteins. Some of the assigned large single-domain proteins may have two or more domains in the actual folding.

Multiple domains in large proteins

Analysis based on known protein structures shows that a peptide chain larger than 250 amino acids barely folds cooperatively into just one domain [23]. Multidomain proteins form through the 'docking' of domains, which are the compact, hydrophobic, independently folding nuclei [12] during the protein folding process. The distribution of the chain lengths in single-domain proteins is similar to the distribution of the number of amino acids in domains of multidomain proteins, (Figure 3a). The periodicity observed in protein sizes [5,6] reflects the distribution of domain sizes. The period of chain length, according to the distribution in Figure 3b, should be between the optimal n value for maximum $P(n)$ ($n_{dom} = 106$) and the average domain size ($n_{ave} = 148$). This is in agreement with the studies in [5,6]. The thermodynamic origin for large proteins to divide into several domains in folding is revealed in our energy functions. The entropic cost for protein folding increases quadratically with increasing chain length in a single-domain protein. By dividing the protein into several domains, the sum of the entropic cost of each domain is substantially less than the cost in forming a single-domain protein.

Some limitations

Our model provides a new perspective for further studies on the structure–energetics relationship of proteins, in particular on the thermodynamic origin of protein sizes. The conceptual framework of our study does not depend on the details of the energy functions. Nevertheless, there are some limitations in our studies due to the limited size and quality of the thermodynamic data. Furthermore, different methods vary in partitioning the free energy of unfolding, especially for the assessment of the configurational entropy [24–28]. Although our energy functions fit the data well, the model is based on statistics rather than on an *ab initio* method. Hence, the functional forms of energetics can be improved from further studies. As more data become available, the new energy functions may result in more accurate estimates of the domain size and enable studies of additional aspects of protein sizes, such as the relationship between the size and the composition of a protein, its shape, or its disulfide bonds.

Acknowledgements

We thank Charles Lawrence, Dong Xie, Richard Goldstein, Eugene Kolker, Chung-Jung Tsai, Shuo L. Lin, Aijun Li, and, in particular, Jacob Maizel for helpful discussions. We thank the personnel at the Frederick Cancer Research and Development Center for their assistance. The calculations presented in this paper were carried out on Silicon Graphics workstations operated by the Frederick Biomedical Supercomputing Center, National Cancer Institute. The research of R.N. has been sponsored by the National Cancer Institute, DHHS, under contract number 1-CO-74102 with SAIC, and in part by grant number 95-00208 from the BSF, Israel, by a grant from the Israel Science Foundation administered by the Israel Academy of Sciences, and by the Rekanati Fund.

References

1. Wetlaufer, D.B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA* **70**, 697-701.
2. Schulz, G.E. & Schirmer, R.H. (1979). *Principles of Protein Structure*. Springer-Verlag, Heidelberg, Germany.
3. Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**, 167-339.
4. Janin, J. & Wodak, S.J. (1983). Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Molec. Biol.* **42**, 21-78.
5. Savageau, M.A. (1986). Proteins of *Escherichia coli* come in sizes that are multiples of 14 kDa: domain concepts and evolutionary implications. *Proc. Natl Acad. Sci. USA* **83**, 1198-1202.
6. Berman, A.L., Kolker, E. & Trifonov, E.N. (1994). Underlying order in protein sequence organization. *Proc. Natl Acad. Sci. USA* **91**, 4044-4047.
7. Russell, R.B. (1994). Domain insertion. *Protein Eng.* **7**, 1407-1410.
8. Tang, J., James, M.N., Hsu, I.N., Jenkins, J.A. & Blundell, T.L. (1978). Structural evidence for gene duplication in the evolution of the acid proteases. *Nature* **271**, 618-621.
9. Lin, X.L., Lin, Y.Z., Koelsch, G., Gustchina, A., Wlodawer, A. & Tang, J. (1992). Enzymic activities of two-chain pepsinogen, two-chain pepsin, and the amino-terminal lobe of pepsinogen. *J. Biol. Chem.* **267**, 17257-17263.
10. Shirai, T. & Go, M. (1997). Adaptive amino acid replacement accompanied by domain fusion in reverse transcriptase. *J. Mol. Evol.* **44**, S155-S162.
11. Wu, L.C., Grandori, R. & Carey, J. (1994). Autonomous subdomains in protein folding. *Protein Sci.* **3**, 359-371.
12. Tsai, C.J. & Nussinov, R. (1997). Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci.* **6**, 24-42.
13. Novokhatny, V.V., Kudinov, S.A. & Privalov, P.L. (1984). Domains in human plasminogen. *J. Mol. Biol.* **179**, 215-232.
14. Shore, D., Langowski, J. & Baldwin, R.L. (1981). DNA flexibility studied by covalent closure of short fragments into circles. *Proc. Natl Acad. Sci. USA* **78**, 4833-4837.
15. Dill, K.A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry* **24**, 1501-1509.
16. Makhatadze, G.I. & Privalov, P.L. (1994). Hydration effects in protein unfolding. *Biophys. Chem.* **51**, 291-309.
17. Miller, S., Janin, J., Lesk, A.M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641-656.
18. Creighton, T.E. (1993). *Proteins.* (2nd edn), W.H. Freeman and Company, New York.
19. Islam, S.A., Luo, J. & Sternberg, M.J. (1985). Identification and analysis of domains in proteins. *Biochemistry* **24**, 1501-1509.
20. Xie, D. & Freire, E. (1994). Structure based prediction of protein folding intermediates. *J. Mol. Biol.* **242**, 62-80.
21. Branden, C. & Tooze, J. (1991). *Introduction to Protein Structure.* Garland Publishing, Inc., New York.
22. Efimov, A.V. (1993). Pattern of loop regions in proteins. *Curr. Opin. Struct. Biol.* **3**, 379-384.
23. Garel, J.R. (1992) In *Protein Folding.* Folding of large proteins: multidomain and multisubunit proteins. (Creighton, T.E., ed.), pp. 405-454, W.H. Freeman and Company, New York.
24. Lazaridis, T., Archontis, G. & Karplus, M. (1995). Enthalpic contribution to protein stability: insights from atom-based calculations and statistical mechanics. *Adv. Protein Chem.* **47**, 231-306.
25. Stites, W.E. & Pranata, J.C. (1995). Empirical evaluation of the influence of side chains on the conformational entropy of the polypeptide backbone. *Proteins* **22**, 132-140.
26. D'Aquino, J.A., Gomez, J., Hilser, V.J., Lee, K.H., Amzel, L.M. & Freire, E. (1996). The magnitude of the backbone conformational entropy change in protein folding. *Proteins* **25**, 143-156.
27. Jelesarov, I. & Bosshard, H.R. (1996). Thermodynamic characterization of the coupled folding and association of heterodimeric coiled coils (leucine zippers). *J. Mol. Biol.* **263**, 344-358.
28. Zhang, C., Cornette, J.L. & Delisi, C. (1997). Consistency in structural energetics of protein folding and peptide recognition. *Protein Sci.* **6**, 1057-1064.