Protein Science (2000), *9*:580–586. Cambridge University Press. Printed in the USA.

Optimal region of average side-chain entropy for fast protein folding

OXANA V. GALZITSKAYA,[1,2] ALEXEI K. SURIN,[3,2] AND HARUKI NAKAMURA[1,4]

[1]Biomolecular Engineering Research Institute, 6-2-3 Furuedai, Suita Osaka 565-0874, Japan
[2]Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russia
[3]Kansai Medical University, 18-89 Uyamahigashi, Hirakata 573-1136, Japan

(RECEIVED August 16, 1999; FINAL REVISION January 3, 2000; ACCEPTED January 9, 2000)

Abstract

Search and study the general principles that govern kinetics and thermodynamics of protein folding generates new insight into the factors that control this process. Here, we demonstrate based on the known experimental data and using theoretical modeling of protein folding that side-chain entropy is one of the general determinants of protein folding. We show for proteins belonging to the same structural family that there exists an optimal relationship between the average side-chain entropy and the average number of contacts per residue for fast folding kinetics. Analysis of side-chain entropy for proteins that fold without additional agents demonstrates that there exists an optimal region of average side-chain entropy for fast folding. Deviation of the average side-chain entropy from the optimal region results in an anomalous protein folding process (prions, α-lytic protease, subtilisin, some DNA-binding proteins). Proteins with high or low side-chain entropy would have extended unfolded regions and would require some additional agents for complete folding. Such proteins are common in nature, and their structure properties have biological importance.

Keywords: Monte Carlo simulations; native topology; optimal balance; protein stability; side-chain packing; unfolded regions

The folding of biological macromolecules is one of the main problems of molecular biology and biophysics. Protein molecules have an enormous number of possible conformations, but this does not hinder finding their unique stable three-dimensional (3D) structure within the time much less than necessary for an exhaustive sorting of all their conformations. Understanding the reason for fast folding of 3D structures of biological macromolecules is especially important for the de-novo design of proteins. For instance, for the de-novo design of a protein, it is necessary to know what features of its primary structure define the stability of its 3D structure and what features of the sequences provide for its fast folding.

Some general trends and correlations are beginning to emerge between the structural, thermodynamic, and kinetic properties of proteins (Jackson, 1998; Plaxco et al., 1998b; Shakhnovich, 1998). Despite many theoretical and experimental efforts in this field, there is no consensus as to what factor (size, topology, or stability) is more important and governs protein folding (Jackson, 1998; Plaxco et al., 1998b). There is the enormous diversity in the folding behavior as well for small proteins that fold with simple two-state kinetics as for large proteins that fold with multi-state kinetics.

Reprint requests to: Oxana V. Galzitskaya, Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russia; e-mail: ogalzit@vega.protres.ru.

[4]Present address: Institute of Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan.

The apparent lack of a relationship between stabilities and folding rates of topologically diverse proteins indicates that topology may be a critical determinant of folding kinetics (Plaxco et al., 1998b). But only topology cannot explain the differences in the refolding rates for some proteins sharing the same fold rates (SH3 domain proteins, fibronectin domains, cold shock proteins, proteins belonging to the ferredoxin fold) (Perl et al., 1998; van Nuland et al., 1998; Zerovnik et al., 1998). Although all factors determining folding rates for given proteins are still not understood, these factors must be related to the intrinsic properties of amino acid residues. Now we know from the experimental data that a single point mutation can dramatically alter the structure and/or refolding of a protein (Matouschek et al., 1995; Milla & Sauer, 1995; Furukawa et al., 1996). Besides, the amino acid residue composition of a protein allows us to determine its structural class (α, β, α/β, $\alpha + \beta$) with a high accuracy (Dubchak et al., 1995; Bahar et al., 1997). It has been emphasized that the loss of chain entropy upon protein folding is a rate determining factor (Finkelstein & Badretdinov, 1997a; Hao & Scheraga, 1998; Plaxco et al., 1998b). The search for the factors affecting the protein folding process continues.

In this work, we demonstrate based on the known experimental data and using theoretical modeling of protein folding process that side-chain entropy is very important factor of protein folding. Analysis of side-chain entropy for proteins that fold without additional agents demonstrates the existence of the optimal region of the

average side-chain entropy for fast folding. Deviation of the average side-chain entropy from the optimal region results in an anomalous protein folding process (prions, α-lytic protease, subtilisin, some DNA-binding proteins). Therefore, proteins with high or low side-chain entropy would have extended unfolded regions and would require some additional agents for complete folding. We demonstrate that the differences in the folding rates can be explained as a result of the differences in the balance between conformational entropy and energy of side-chain interactions considering the theoretical model (Galzitskaya & Finkelstein, 1999) of protein folding of structurally similar proteins for three different protein families (SH3 domain family, cold shock DNA-binding domain-like family, and proteins belonging to the ferredoxin fold). Namely, the existence of such balance helps to explain why sometimes the less stable protein folds faster by an order of magnitude than the more stable protein with the same topology (van Nuland et al., 1998).

Results

Analysis of average side-chain entropy

To examine whether folding in proteins has a correlation with average side-chain entropy, we have analyzed proteins taken from a recent review (Jackson, 1998) where their folding processes were shown experimentally have two- and three-state kinetics. The value of the average side-chain entropy ν was calculated as a summation of the individual side-chain entropy of a residue over its complete sequence normalized by the protein length using a side-chain entropy scale developed by Pickett and Sternberg (1993). Figure 1 demonstrates that topologically diverse proteins that fold without additional agents fall in a definable region of the average side-chain entropy (filled circles). Proteins with higher or lower side-chain entropy (open circles) as a rule have extended unfolded regions and/or require some additional agents for complete folding. For example, some DNA-binding proteins such as transcription factors (with high side-chain entropy) have regions which became structured only upon DNA binding (McIntosh et al., 1998; Ippel et al., 1999; Wright & Dyson, 1999); α-lytic protease and subtilisin (with low side-chain entropy because of the relatively high Gly and Ala content that results in a high conformational entropy of the backbone chain (D'Aquino et al., 1996)) use a pro region to promote folding (Baker, 1998; Sohl et al., 1998); the N-terminal region of prions (James et al., 1997) with low side-chain entropy can adopt a more stable conformation that is induced by interaction with other prions with that conformation (Taylor et al., 1999); the C-terminal domain of PGK (ν = 1.56, out of the optimal region) refolds 40 times slower than the N-domain (Parker et al., 1996); hemoglobin (ν = 1.56—out of the optimal region) has an oligomeric structure, at the same time myoglobin (ν = 1.74—inside the optimal region) has a monomeric form. It seems that proteins with a large conformational entropy (out of the optimal region) have no sufficient energetic interactions to compensate such large entropy. Therefore, enhanced stabilization for them is achieved by additional interactions with other agents or by oligomerization.

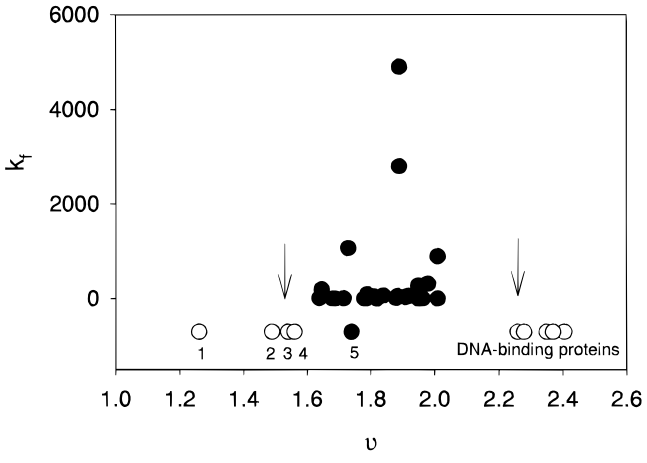

Fig. 1. Dependence of the protein folding rate k_f on the average side-chain entropy ν. Filled circles correspond to the proteins that fold with two- and three-state kinetics (Jackson, 1998): PDB files are *2abd*, *1hcr*, *1csp*, *1a0n*, *2ci2*, *1hdn*, *2ptl*, *1aps*, *1pba*, *1ten*, *1lmb*, *1mjc*, *1shg*, *1srl*, *1pks*, *2ait*, *1urn*, *1fkb*, *1bta*, *1bni*, *1hel*, *1ubq*, *3chy*, *2rn2*, *1php*. Open circles correspond to the proteins that have peculiarities in the folding process. Right arrow corresponds to the DNA-binding proteins (PDB files are *1fjl*, *1ftz*, *1hdp*, *1pog*, *2hoa*). Left arrow corresponds to: 1, the N-terminal extended region of prion; 2, subtilisin; 3, α-lytic protease; 4, the C-terminal domain of PGK and hemoglobin; 5, myoglobin.

Monte Carlo simulations of protein folding

To demonstrate the influence of the side-chain entropy on the rate of folding and importance of the balance between the side-chain entropy and energy of residue–residue interactions, the theoretical model of protein folding following from the analytical theory of this process (Finkelstein & Badretdinov, 1997a, 1997b) was considered. We performed traveling from the unfolded state to the native 3D structure without misfolding to other compact states. In this model, the folding pathways are treated as sequential insertion of residues from the unfolded state to their native positions according to the 3D native structure or removal of residues from the native position to the coil, respectively (Fig. 2).

The removed (inserted) residues are assumed to loose (gain) all the nonbonded interactions and gain (loose) the coil entropy except that spent to close the disordered loops protruding from the remaining globule (Galzitskaya & Finkelstein, 1999). The general assumption of this model is that the residues remaining in the globule keep their native position and that the unfolded regions do not fold to another, nonnative globule. Thus, we neglect nonnative interactions that make our model similar to that of Go (Ueda et al., 1975).

We consider our model as an approximation of the protein folding process, rather than a detailed description of the chain motions. Thereby our model of folding is a trade-off between the configurational entropy loss and the gain of attractive interactions. In such model of protein folding, the folding rate decreases with increasing temperature (Baker, 1998). Therefore, we made our simulations at low temperatures (RT = 0.6, energy units). The model takes into account the topology of the native state and the side-chain conformational entropy. Such consideration is important to describe and explore additional properties of protein folding that may be a consequence of various sizes, degrees of freedom, and shapes of amino acid residues.

To investigate the influence of the side-chain entropy on the time of folding, three protein sets (SH3-domain family, cold shock DNA-binding family, and proteins belonging to the ferredoxin fold)

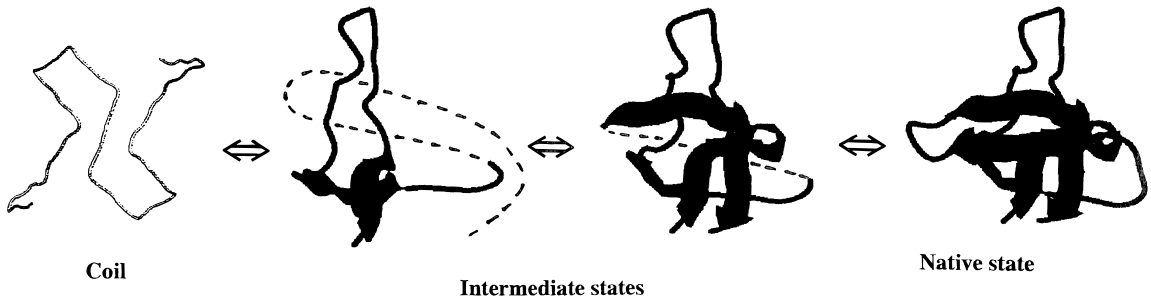

Fig. 2. Schematic presentation of a protein folding pathway. Residues in the intermediate states keep their native positions (solid line), while other residues are unfolded (broken line).

with opposite dependencies between folding and stability were used (see Materials and methods). Monte Carlo simulations were done to calculate how long a given protein chain folds to the native structure.

Figure 3 shows the dependence of the characteristic first passage time $t_{1/2}$ measured as the number of Monte Carlo steps on the average side-chain entropy for the three sets of proteins. It should be noted that the average side-chain entropy larger than two is absent for these families. Such increase of this value could lead to a fatal decrease of the protein stability and such proteins have been eliminated during the protein evolution process of these families. The dependencies demonstrate that the proteins with some intermediate values of the average side-chain entropy (about 1.7–1.9 entropy units divided by R, R is the gas constant) fold faster than other ones. To clarify this fact, the available experimental data for these families (Guijarro et al., 1998; Jackson, 1998; Perl et al., 1998; Plaxco et al., 1998a, van Nuland et al., 1998) are presented. Figure 4 shows the dependencies of the inverse constant of folding on the average side-chain entropy. Such dependencies also underline the existence of the optimal region of the average side-chain entropy for fast folding.

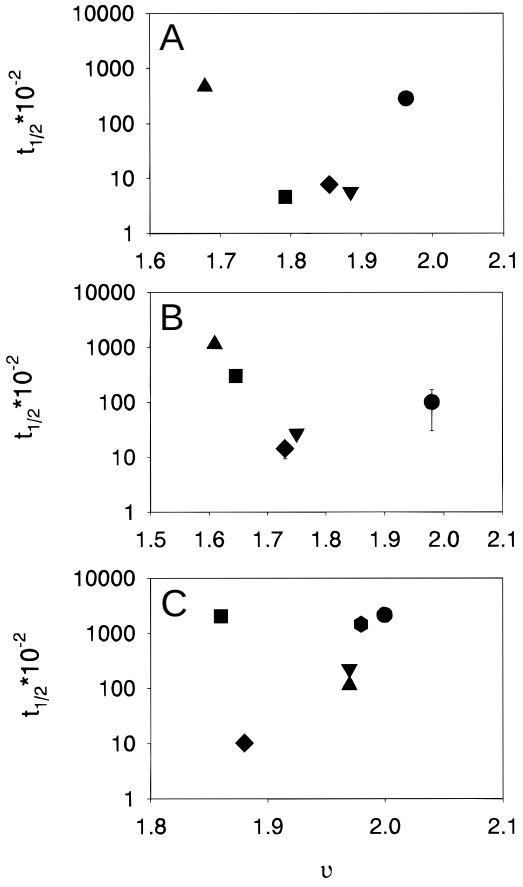

Fig. 3. Dependence of the characteristic first passage time $t_{1/2}$ on the average side-chain entropy ν (entropy units divided by R). All simulations have been done at fixed temperature $RT = 0.6$. **A:** For the SH3-domain family: *1shg* (●), *1srm* (▼), *1nyf* (■), *1cka* (♦), and *1gfc* (▲). **B:** For the cold shock family: *1ah9* (●), *1sro* (▼), *1mjc* (■), *1nmg* (♦), and one β protein, *2ait* (▲). **C:** For the ferredoxin-like fold: *1aps* (●), *1urn* (▼), *1sxl* (■), *1hdn* (♦), *1ris* (▲), and *2acy* (♦). Errors are shown by vertical bars when they exceed the symbol size.

Different values of the optimal side-chain entropy in the kinetic experiments for the different families can be a result of the specificity in the packing of side chains depending on the type of fold (Behe et al., 1991). To obtain more information, we also checked the influence of the number of contacts on the folding process. The calculations were done without consideration of the side-chain entropy. Figure 5 shows the absence of a distinct correlation between the folding rate and the average number of contacts per residue. It is noteworthy that proteins with the smallest value of the average side-chain entropy have the largest value of the average number of contacts.

"Entropy capacity" for protein chain with given topology

The formation of sufficient residue–residue interactions is necessary to compensate the side-chain conformational entropy during the protein folding process. Therefore, structural uniqueness of native proteins is the result of the balance between the conformational entropy of side chains and the energy of residue interactions. Taking these phenomena into account, we introduced a new parameter—"entropy capacity"—for a given protein chain with a chosen topology. This parameter means the relation between side-chain entropy and residue–residue interactions that give a large contribution to the free energy of the system. The relationship between these values will determine the possibility of the given chain to fold to that particular topology.

The nucleation barrier of protein folding is intrinsic for the folding process as each protein first has to form its nucleus. Usually, this barrier is much smaller than the barrier between the molten globule and the native states (Ptitsyn, 1995). But the second barrier involving the tight packing of side chains can be dra-

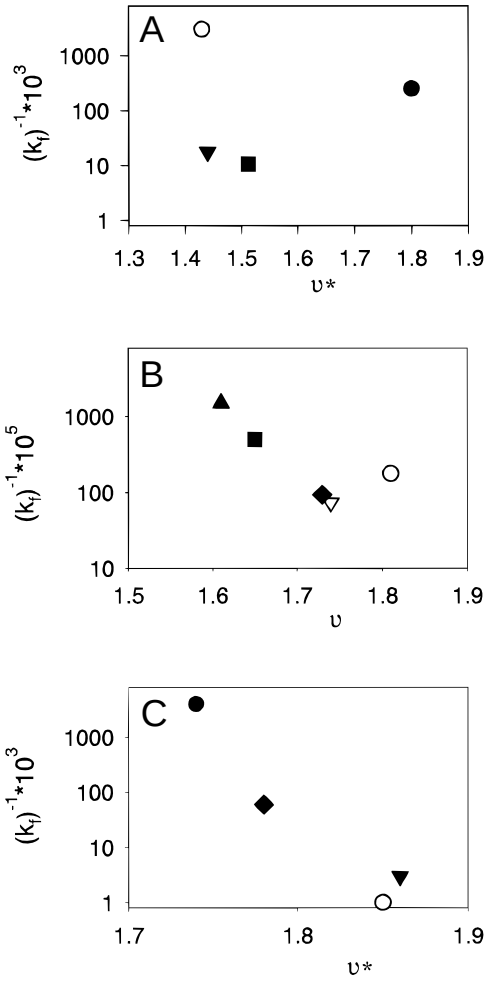

Fig. 4. Dependence of the inverse constant of folding $(k_f)^{-1}$ on the average side-chain entropy for all residues ν and for buried side-chains ν*. The latter is defined by the criterion that the accessible surface area of a residue is less than 60% of the value for the residue in the extended state (Pickett & Sternberg, 1993). **A:** SH3-domain family: *1shg* (●), *1srm* (▼), *1nyf* (■), 1pks (○). **B:** Cold shock family: *1nmg* (♦) *CspB* (*Bacillus subtilis*), *CspB* from *Bacillus caldolyticus* (▽), *CspB* from *Thermotoga maritima* (○), *1mjc* (■), *2ait* (▲). **C:** Ferredoxin fold: *1aps* (●), *1urn* (▼), *1hdn* (♦), *1pba* (○).

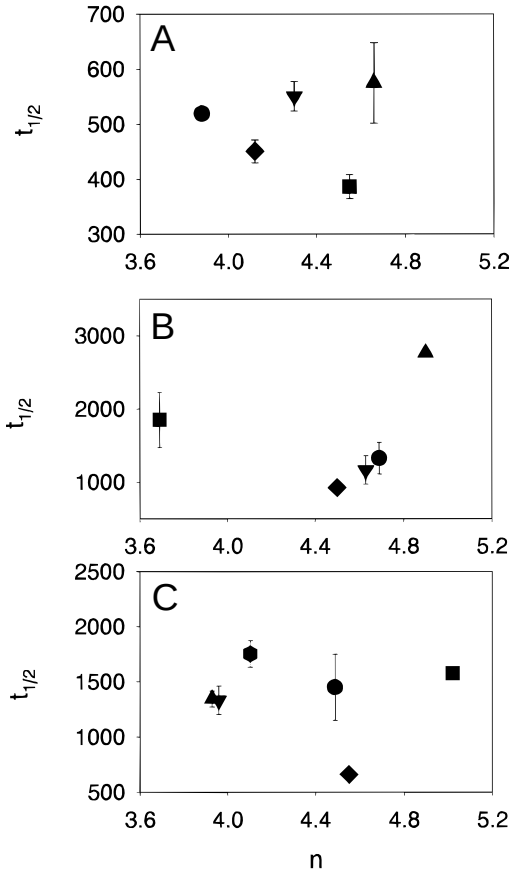

Fig. 5. Characteristic first passage folding time $t_{1/2}$ on the average number of contacts per residue. All simulations without consideration of the side-chain entropy have been done at fixed temperature $RT = 0.6$. This figure is depicted as described in Figure 3.

matically decreased and in these cases the nucleation barrier represents a rate-limiting step of protein folding. Considering the second stage of protein folding procedure, we can estimate the changing of free energy of this process for the capillarity model (Finkelstein & Badretdinov, 1997a; Wolynes, 1997) as

$$\Delta F = \epsilon(m - \mu m^{2/3}) - \nu m RT. \quad (1)$$

Here, m is the number of fixed residues in the native topology, ϵ is the average contact energy per residue, $\mu m^{2/3}$ takes into account the surface residues having less interactions than the internal ones, where $\mu = 1.5$ for a ball-like body (Finkelstein & Badretdinov, 1997a), ν is the average side-chain entropy per residue for an unfolded chain, and T is the temperature. Then we can estimate the optimal number of fixed residues at a maximum value of ΔF:

$$m_{opt} = (1 - C)^{-3}, \quad (2)$$

where $C = \nu RT/\epsilon$ is the new parameter, i.e., entropy capacity. Therefore, this parameter determines the optimal number of fixed residues m_{opt} (residues that have at least one native contact for the given topology) in a hydrophobic core for a domain with a unique structure.

Figure 6 shows the dependence of the optimal number of fixed residues in the native topology on the entropy capacity. In other words, this figure reflects the existence of the balance between the conformational entropy of side chains and the energy of residue–residue interactions. An increase in the latter leads to a loss of unique structure (a decrease in the number of native contacts—small value of m_{opt}) because the protein can adopt different conformations. While both the decrease in energy and/or any deviations of the average side-chain entropy from the optimal region result in a decrease of protein stability up to its full loss (an increase in the number of contacts over the native ones—large value of m_{opt}). So that the value of m_{opt} would correspond to the real number of the native contacts, there must be some optimal value of entropy capacity to realize a stable unique structure for the whole protein molecule and achievement of the native state.

Figure 7 demonstrates the existence of the optimal region for the new parameter, entropy capacity, to enable fast folding of the same set of proteins. In our theoretical model, this parameter has a simple determination $C_{mod} = \nu/n$, where n is the number of residue–

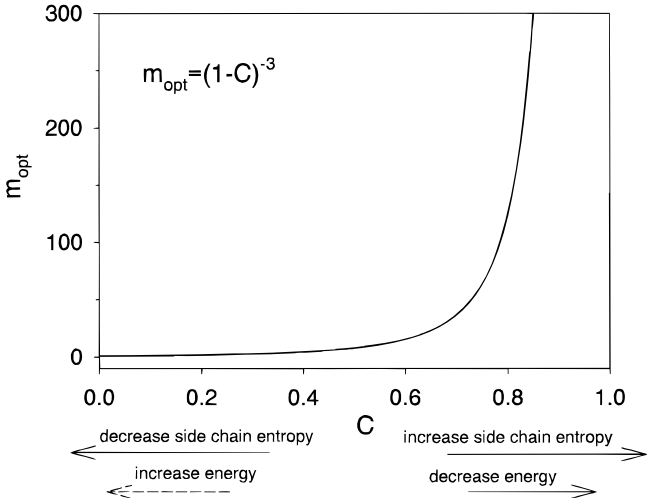

Fig. 6. Dependence of the optimal number of native fixed residues m_{opt} on the entropy capacity C. Entropy capacity depends on the amino acid sequence under the given topology. Arrows indicate effects from the factors that cause loss in stability (solid line) or loss of uniqueness (broken line).

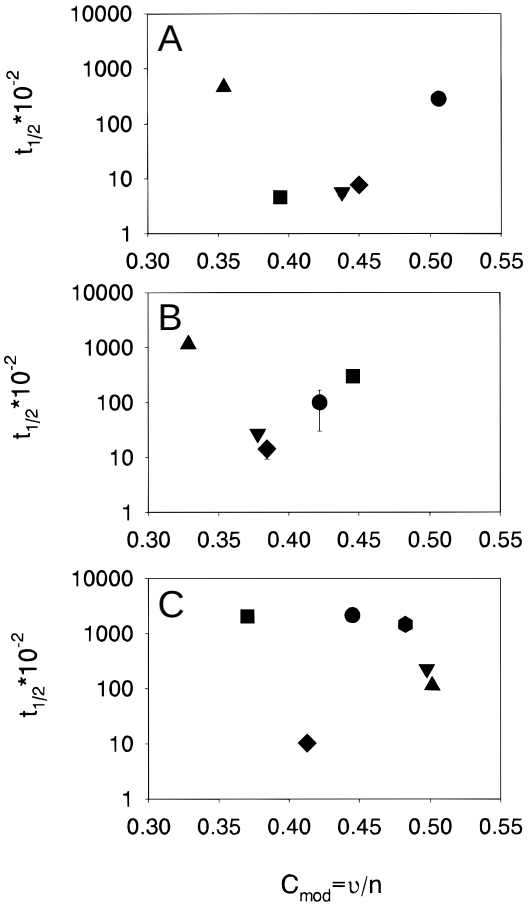

Fig. 7. Dependence of the characteristic first passage folding time $t_{1/2}$ on the model entropy capacity $C_{mod} = \nu/n$. This figure is depicted as described in Figure 3.

residue contacts in the native structure and ν is the side-chain entropy. Scaled by n, C_{mod} has a common optimum value about 0.4–0.5 corresponding to fast folding for the three sets of proteins.

Discussion

The presented model of protein folding is a rough approximation of the process. But it can describe a substantial point of the folding procedure, even if the actual value of some parameters deviate from the experimental and simulation values. This model (only without side-chain entropy consideration) has been employed to predict the structure of folding nuclei in 3D protein structures using dynamic programming (Galzitskaya & Finkelstein, 1999). This model neglects ruggedness of the landscape. But the fluctuation effects can lead to a very broad trapping-time distribution (Wolynes, 1997). Consideration of models with native topology and side-chain conformations of side-chain groups is an efficient way to explore global structural features of biological molecules and provide some information about the folding pathway both in theoretical models and in real proteins.

The presented theoretical simulations and the available experimental data for the rate of folding show the existence of the optimal region of side-chain entropy for fast folding in the every considered set of proteins sharing the same fold. This underlines the important role of the side-chain entropy in the protein folding process. In our theoretical modeling of protein folding, the rate of achievement of the native state is limited by insufficient thermodynamic stability of proteins at high values of the average side-chain entropy, while at low values of the average side-chain entropy it is limited by a necessity to sort out conformations (with several residues not fixed in the native position) more stable than the native one.

We did not take into account the contribution of the solvent in our model. But the influence of this factor has been considered in the presentation of experimental data for the rate of folding. The obtained dependencies (Fig. 4A,C) of the inverse constant of folding on the average side-chain entropy only for residues that belong to the hydrophobic core also demonstrate the existence of the optimal region of the average side-chain entropy for fast folding.

The existence of the optimal region of the entropy capacity relating two important factors of protein folding such as the side-chain entropy and the energy of residue–residue interactions suggests some optimum balance between them for fast folding. Namely, taking into account the existence of such optimum, we can explain why sometimes the less stable proteins fold faster than the more stable proteins with the same topology (Perl et al., 1998; van Nuland et al., 1998; Zerovnik et al., 1998).

The need to obtain a definite balance between the conformational entropy and the energy of interactions is one of the general conditions to achieve the functional active form of the protein. For some proteins (with high conformational entropy or low energy of residue–residue interactions), such balance can be achieved only by oligomerization with the same proteins and for other proteins only by interaction with additional agents (Wright & Dyson, 1999).

The shift of this balance due to the changing of the external condition (pH or temperature) can result in the formation of stable intermediates such as molten globule-like intermediates (Ptitsyn, 1995) or fibrils that may play a pathological role in the cell (Chiti et al., 1999; Harrison et al., 1999; Jimenez et al., 1999).

Considering the theoretical dependence of the folding rate on the both parameters (average side-chain entropy and entropy ca-

pacity), we can estimate which protein from a family is the most representative for a given fold. Since the protein with the optimal relationship between energetic and entropic parameters would show fast kinetics and sufficient thermodynamic stability.

The obtained results may be useful in protein design. When we attempt to create a de-novo protein that has a unique structure, it is important to consider the side-chain conformational entropy that is certain to belong to the definite region of the average side-chain entropy scale.

Materials and methods

Free energy estimate

Considering the pathway of folding to the native 3D structure, the free energy of intermediate structures with different number of fixed residues in the native position can be calculated according to Equation (3):

$$F(B_k) = E(B_k) - S(B_k)\cdot T = \epsilon \cdot \sum_{i,\, j\in B_k} \sum_{(i<j)} \delta_{ij} - RT \times \sum_{i\notin B_k} \sigma_i - T\cdot \sum_{loops\in B_k} S_{loop} - RT\cdot \sum_{i\notin B_k} \nu_i \qquad (3)$$

where k residues are unfolded in the intermediate structure B_k, while the other residues keep their native positions (Galzitskaya & Finkelstein, 1999). $E(B_k)$ is proportional to the number of contacts in the intermediate structure B_k. Besides, it has been shown experimentally that the number of van der Waals contacts turns out to be more suitable for analyzing structural and energetic responses to mutation than other parameters (Vlassi et al., 1999).

The first summation is taken over all nonneighbor residues ($i < j$) keeping their native positions in B_k. $\delta_{ij} = 0$ if i and j residues have no contacts and $\delta_{ij} = 1$ if such a contact is present. Two residues are assumed to be in contact when the minimal distance between their atoms is <5 Å. $\epsilon = -1$, where 1 is the energy unit of one residue–residue contact. Conformational entropy has been subdivided into backbone and side-chain contributions.

In the second summation, σ_i is the backbone entropy difference between the coil and the native state of a residue. From review of Brady and Sharp (1997), we used $\sigma_i = 2.3$ (entropy units divided by R, R being the gas constant) for all residues with exception of Gly, $\sigma_i = 3.2$, and Pro, $\sigma_i = 0$.

The third summation is taken over all closed loops protruding from the globular part of B_k. The entropy spent to close a disordered loop between fixed residues m and l is estimated (Finkelstein & Badretdinov, 1997b) as

$$S_{loop} = -5/2\ R \ln|m-l| - 3/2\ R(r_{ml}^2 - a^2)/(2Aa|m-l|) \quad (4)$$

where r_{ml} is the distance between the C_α atoms of the residues m and l, $a = 3.8$ Å is the distance between the neighbor C_α atoms in the chain, and A is the persistent length for a polypeptide (according to Flory, 1969, $A = 20$ Å).

The last summation corresponds to the side-chain entropy of unfolded residues. ν_i is the side-chain entropy for the i^{th} unfolded residue. We used side-chain entropy parameters developed by Pickett and Sternberg (1993). This side-chain entropy scale correlates well with other scales (Doig & Sternberg, 1995). We assumed that the side-chain entropy in the folded state is equal to zero.

Investigation of folding kinetics

We calculated how long a given protein chain folds to the native structure, starting from six arbitrarily chosen fixed residues (this is a minimal number of native residue positions to shift the equilibrium for the native structure formation under considered temperature) by Monte Carlo (MC) simulation using the Metropolis scheme (Metropolis et al., 1953) at low temperatures ($RT = 0.6$ energy units). To shift the equilibrium in the MC simulations without consideration of side-chain entropy, it is sufficient to fix four arbitrary residues. The kinematic scheme of elementary movements includes removal of a residue from the native position to the coil or insertion of a residue from the coil to the native position (Galzitskaya & Finkelstein, 1998). We did traveling from the unfolded state to the native 3D structure without misfolding to other compact states.

An elementary MC step was done as follows. We randomly chose a residue. If the chosen residue had been already fixed in the native position, we tried to unfold it. If the chosen residue was in the coil, we tried to fix it according to its native position. Then we computed the free energy difference ΔF between new and previous intermediate structures. The MC step leads to the new structure with a probability w, which is equal to $\exp(-\Delta F/RT)$, if $\Delta F > 0$, or to 1, if $\Delta F \le 0$. Thus, if $\Delta F \le 0$, the MC step leads to the new structure automatically. If $\Delta F > 0$, w is compared with a random number ξ ($0 < \xi < 1$; ξ is generated according to the uniform distribution) if $w > \xi$, the transition is accepted; if $w \le \xi$, it is rejected and the previous state is preserved.

To estimate the characteristic first passage time $t_{1/2}$ and an error of the simulation for a given protein chain and given temperature, we performed two sets of 50 MC runs (Galzitskaya & Finkelstein, 1998). For every set of values, $t_{1/2}$ was determined as the number of MC steps required to complete 50% of MC runs (25 of 50 runs). Half-summation of these values ($t'_{1/2}$ and $t''_{1/2}$) gives the estimation of the average characteristic of the first passage time and their half-difference gives the characteristic value of error in the estimation of this time:

$$t_{1/2} = t^0_{1/2} \pm \delta t_{1/2} = (t'_{1/2} + t''_{1/2})/2 \pm |t'_{1/2} - t''_{1/2}|/2. \qquad (5)$$

Protein families

For theoretical modeling of folding, three protein sets with opposite dependencies between folding and stability were used. The proteins of a similar length, especially from the same topological family taken from SCOP (Murzin et al., 1995) and with known experimental kinetic data, were chosen. The 3D coordinates of the native structures were taken from the Protein Data Bank (PDB) files (Bernstein et al., 1977).

The first group is SH3-domain family where the most stable protein has been shown experimentally to fold the most rapidly (Plaxco et al., 1998a). SH3 domains are small, monomeric domains without disulfide bonds and prosthetic groups. Interestingly, SH3 domains with nearly identical backbone conformations have different folding rates ranging from 94 s^{-1} for the *Fyn* domain to 0.35 s^{-1} for the *PI3*-kinase domain at 20 °C (Jackson, 1998). We used five proteins from this family: PDB files are *1shg, 1srm, 1nyf, 1cka, 1gfc*. PI3-kinase has not been considered in the simulation because it has a longer length of the chain than the above-mentioned proteins.

The second group is the cold shock DNA-binding domain-like family where the most stable protein does not fold the most rapidly

(Perl et al., 1998). Cold shock protein *B*, *CspB*, has a β-barrel fold and is one of the fastest folding proteins (1,070 s^{-1} at 25 °C). Besides, *CspB* has the most highly structured transition states than other proteins studied to date (Schindler et al., 1995). We have used four proteins from this family: *1ah9*, *1sro*, *1mjc*, *1nmg*, and one β-protein—*2ait*.

The third group is the ferredoxin-like fold where proteins have a classical open-faced β-sandwich fold comprised of two or three antiparallel α-helices packed against a β-sheet. Acylphosphatase (*1aps*) folds more than 1,000-fold slower than the activation domain procarboxypeptidase A2 (*1pba*). The latter with the lowest stability refolds the most rapidly (van Nuland et al., 1998). We used the following proteins: *1aps*, *1urn*, *1sxl*, *1hdn*, *1ris*, *2acy*. Procarboxypeptidase A2 (*1pba*) has not been considered in the simulation because it has not enough compact structure in the PDB file. So we can not observe a full folding process for it.

All proteins from these families provide a simple model for the study of protein folding.

Acknowledgments

We are grateful to Profs. A.V. Finkelstein and H. Kihara, and Drs. S. Tsutakawa, J. Higo, and A.A. Timchenko for comments and discussions. A.K. Surin is supported by Japan Society for the Promotion of Science.

References

Bahar I, Atilgan AR, Jernigan RL, Erman B. 1997. Understanding the recognition of protein structural classes by amino acid composition. *Proteins 29*:172–185.

Baker D. 1998. Metastable states and folding free energy barriers. *Nat Struct Biol 5*:1021–1024.

Behe MJ, Lattman EE, Rose GD. 1991. The protein-folding problem: the native fold determines packing, but does packing determine the native fold? *Proc Natl Acad Sci USA 88*:4195–4199.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem 80*:319–324.

Brady P, Sharp KA. 1997. Entropy in protein folding and in protein-protein interaction. *Curr Opin Struct Biol 7*:215–221.

Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G, Dobson CM. 1999. Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc Natl Acad Sci USA 96*:3590–3594.

D'Aquino JA, Gomez J, Hilser VJ, Lee KH, Amzel LM, Freire E. 1996. The magnitude of the backbone conformational entropy change in protein folding. *Proteins 25*:143–156.

Doig AJ, Sternberg JE. 1995. Side-chain conformational entropy in protein folding. *Protein Sci 4*:2247–2251.

Dubchak I, Muchnik I, Holbrook SR, Kim SH. 1995. Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci USA 92*:8700–8704.

Finkelstein AV, Badretdinov AYa. 1997a. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold Des 2*:115–121.

Finkelstein AV, Badretdinov AYa. 1997b. Physical reasons for fast folding of stable spatial structure of proteins: A solution of the Levinthal paradox. *Mol Biol (Russia, Engl. Edition) 31*:391–398.

Flory PJ. 1969. *Statistical mechanics of chain molecules*. New York: Interscience.

Furukawa K, Oda M, Nakamura H. 1996. A small engineered protein lacks structural uniqueness by increasing the side-chain conformational entropy. *Proc Natl Acad Sci USA 93*:13583–13588.

Galzitskaya OV, Finkelstein AV. 1998. Folding rate dependence on the chain length of RNA-like heteropolymers. *Fold Des 3*:69–78.

Galzitskaya OV, Finkelstein AV. 1999. A theoretical search for folding/unfolding nuclei in 3D protein structures. *Proc Natl Acad Sci USA 96*:11299–11304.

Guijarro JI, Morton CJ, Plaxco KW, Campbell ID, Dobson CM. 1998. Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *J Mol Biol 276*:657–667.

Hao MH, Scheraga HA. 1998. Molecular mechanisms for cooperative folding of proteins. *J Mol Biol 277*:973–983.

Harrison PM, Chan HS, Prusiner SB, Cohen FE. 1999. Thermodynamics of model prions and its implications for the problem of prion protein folding. *J Mol Biol 286*:593–606.

Ippel H, Larsson G, Behravan G, Zdunek J, Lundqvist M, Schleucher J, Lycksell PO, Wijmenga S. 1999. The solution structure of the homeodomain of the rat insulin-gene enhancer protein isl-1. Comparison with other homeodomains. *J Mol Biol 288*:689–703.

Jackson SE. 1998. How do small single-domain proteins fold? *Fold Des 3*:R81–R91.

James TL, Liu H, Ulyanov NB, Farr-Jones S, Zhang H, Donne DG, Kaneko K, Groth D, Mehlhorn I, Prusiner SB, Cohen FE. 1997. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc Natl Acad Sci USA 94*:10086–10091.

Jimenez JL, Guijarro JI, Orlova E, Zurdo J, Dobson CM, Sunde M, Saibil HR. 1999. Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing. *EMBO J 18*:815–821.

Matouschek A, Otzen DE, Itzhaki LS, Jackson SE, Fersht AR. 1995. Movement of the position of the transition state in protein folding. *Biochemistry 34*:13656–13662.

McIntosh PB, Frenkiel TA, Wollborn U, McCormick JE, Klempnauer KH, Feeney J, Carr MD. 1998. Solution structure of the B-Myb DNA-binding domain: A possible role for conformational instability of the protein in DNA binding and control of gene expression. *Biochemistry 37*:9619–9629.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys 21*:1087–1092.

Milla ME, Sauer RT. 1995. Critical side-chain interactions at a subunit interface in the Arc repressor dimer. *Biochemistry 34*:3344–3351.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247*:536–540.

Parker MJ, Spencer J, Jackson GS, Burston SG, Hosszu LL, Craven CJ, Waltho JP, Clarke AR. 1996. Domain behavior during the folding of a thermostable phosphoglycerate kinase. *Biochemistry 35*:15740–15752.

Perl D, Welker CH, Schindler TH, Schroder K, Marahiel MA, Jaenicke R, Schmid FX. 1998. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat Struct Biol 5*:229–235.

Pickett SD, Sternberg MJ. 1993. Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol 231*:825–839.

Plaxco KW, Guijarro JI, Morton CJ, Pitkeathly M, Campbell ID, Dobson CM. 1998a. The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry 37*:2529–2537.

Plaxco KW, Simons KW, Baker D. 1998b. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol 277*:985–994.

Ptitsyn OB. 1995. Molten globule and protein folding. *Adv Protein Chem 47*:83–229.

Schindler T, Herrier M, Marahiel MA, Schmid FX. 1995. Extremely rapid protein folding in the absence of intermediates. *Nat Struct Biol 2*:663–673.

Shakhnovich EI. 1998. Protein design: A perspective from simple tractable models. *Fold Des 3*:R45–R58.

Sohl JL, Jaswal SS, Agard DA. 1998. Unfolded conformations of α-lytic protease are more stable than its native state. *Nature 395*:817–819.

Taylor KL, Cheng N, Williams RW, Steven AC, Wickner RB. 1999. Prion domain initiation of amyloid formation in vitro from native Ure2p. *Science 283*:1339–1343.

Ueda Y, Taketomi H, Go N. 1975. Studies on protein folding, unfolding and fluctuations by computer simulation. *Int J Peptide Protein Res 7*:445–459.

van Nuland NAJ, Chiti F, Taddei N, Raugei G, Ramponi G, Dobson CM. 1998. Slow folding of muscle acylphosphatase in the absence of intermediates. *J Mol Biol 283*:883–891.

Vlassi M, Cesareni G, Kokkinidis M. 1999. A correlation between the loss of hydrophobic core packing interactions and protein stability. *J Mol Biol 285*:817–827.

Wolynes PG. 1997. Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc Natl Acad Sci USA 94*:6170–6175.

Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J Mol Biol 293*:321–331.

Zerovnik E, Virden R, Jerala R, Turk V, Waltho JP. 1998. On the mechanism of human stefin B folding: I. Comparison to homologous stefin A. Influence of pH and trifluoroethanol on the fast and slow folding phases. *Proteins 32*:296–303.