Pacific Symposium on Biocomputing 5:128-139 (2000)

FOLDING NUCLEI IN 3D PROTEIN STRUCTURES

O.V. GALZITSKAYA
Biomolecular Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565-0874, Japan (on leave from the Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142292 Russia)

A.V. SKOOGAREV, D.N. IVANKOV, A.V. FINKELSTEIN
Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142292 Russia

This paper presents and analyzes the results of several new approaches to the problem of finding the folding nucleus in a given 3D protein structure. Firstly, we show that the participation of residues in the hydrophobic core and the secondary structure of native protein has a rather modest correlation with the experimentally found Φ values characterizing the participation of residues in the folding nuclei. Then we tried to find the nuclei as the free energy saddle points on the network of the folding/unfolding pathways using the branch-and-bound technique and dynamic programming. We also attempted to estimate the Φ values from solving of kinetic equations for the network of protein folding/unfolding pathways. These approaches give a better correlation with experiment, and the estimated folding time is consistent with the experimentally observed rapid folding of small proteins.

1 Introduction

An understanding of the mechanism of protein folding can help in design of new proteins, in understanding of correct and wrong folding of proteins, in attempts to predict protein structure from sequence.

A key event in protein folding is the formation of the folding nucleus [1-4]. This "nucleus" is *un*stable: it corresponds to the transition state (TS), i.e., to the free energy maximum at the folding/unfolding pathway (or, the better to say, to a saddle point at the free energy landscape covered with the network of such pathways).

So far, there is only one, very difficult experimental method to identify the folding nuclei in proteins: to find the residues whose mutations affect the folding rate by changing the TS stability as strongly as that of the native protein [5].

Several approaches have been recently suggested for the theoretical search of folding nuclei in proteins. The first is based on a search for a set of highly conserved residues having no obvious functional role [6,7]; however, this approach can give only a common part of the nuclei existing in homologous proteins. The second approach is based on the correlation between the participation of residues in the folding nucleus and their fluctuations in partly unfolded stationary states [8] or in native proteins [9]. The third, more direct approach is based on all-atom molecular dynamic simulations of protein unfolding [10-13]. However, these simulations need extremely denaturing conditions (500^{o}K, etc.) to be completed in a reasonable time. Therefore the TS found for such extreme unfolding can be rather different from that existing for folding [14].

Here we present and investigate two novel approaches to the search for the folding nuclei. Both suggested approaches are based on the investigation of unfolding pathways of 3D protein structure. The first searches for the TS at these pathways using the branch and bound (BB) technique and the dynamic programming (DP) method. The second approach is based on the solution of kinetic equations for the network of protein folding/unfolding pathways. These approaches give a visible correlation with experiment. (To have a reference point, we previously show that the correlation of the participation of residues in the protein core and in the secondary structure with their involvement in the experimentally found folding nuclei is rather low.)

We investigate protein unfolding rather than folding because this is simpler (since one can avoid exploring numerous high-energy dead-ends of folding), – while, according to the detailed balance principle [15], the pathways for folding and unfolding must coincide when both processes take place under the same conditions. Hence, we are interested in conditions close to that of thermodynamic equilibrium between the native and the coil states. Under these conditions small proteins demonstrate the "all-or-none" transitions, both in thermodynamics [16] and kinetics [1,2]. This allows us to consider only the pathways going from the native to the unfolded state and to neglect those leading to misfolded globules [17].

2. Materials and Methods

2.1 Statistical analysis and correlations

The number of those atomic contacts of residue i of protein p that disappear after unfolding is computed as $C_d(p,i) = \Sigma_k \Sigma_{i<k+1} \delta^d_{ik}$, where δ^d_{ik} is the number of contacts between atoms of residues i and k (atoms are in contact when the distance between them is below d), and the sum is taken over all non-neighbor chain residues of the protein (since the neighbors are in contact at any chain structure).

The involvement of residue i in the hydrophobic core is computed as

$$C^*_d(p,i) = C_d(p,i) / C_d^{\max}(a_i) , \qquad (1)$$

where $C_d^{max}(a_i)$ is the maximal (at a given threshold d) number of contacts for amino acid a_i in all the studied proteins. The secondary structure is determined by program DSSP [18] from the atomic coordinates taken from PDB [19].

The correlation between the experimental $\Phi_f(i)$ values (see below) and any calculated values $A(i)$ is computed (for each protein p) from the conventional equation $Corr_p(A,\Phi_f)=[<A\Phi_f>_p - <A>_p<\Phi_f>_p]/[(<A^2>_p - <A>_p^2)(<\Phi_f^2>_p - <\Phi_f>_p^2)]^{1/2}$, where $<A\Phi_f>_p = (1/N_p) \Sigma_i^{Np} A(i)\Phi_f(i)$, etc.; the sums over i are taken over all the N_p residues of protein p with the experimentally studied Φ_f values. The averaging of the values $Corr_p(A,\Phi_f)$ over all the analyzed proteins is done with the weights proportional to the number N_p of the studied residues in each of the proteins.

2.2 Network of protein unfolding pathways

We consider a network of simplified stepwise unfolding pathways (Fig.1), each step being the removal of one chain link from the native protein 3D structure.

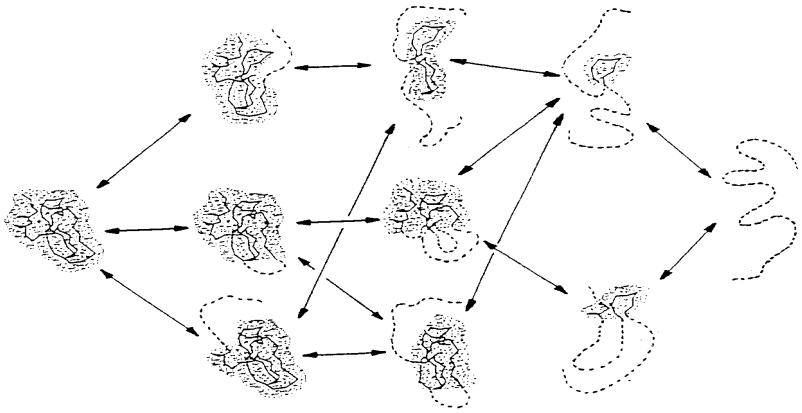

Fig.1. Unfolding intermediates (only a small part of them is shown) and a network of unfolding pathways. Each arrow corresponds to an elementary step, i.e., to the transition of one chain link (of one or a few residues) from the globular, native-like part of the intermediate (bold line), to the coil (dotted line).

The structure S_ν ($\nu = 0,1,..,U$) contains ν disordered and U-ν ordered links; S_0 is the native state, S_U the coil. The removed links are assumed to form the random coil; they lose all the non-bonded interactions and gain the coil entropy (except the entropy spent to close disordered loops protruding from the remaining globule). It is assumed that the links remaining in the globule keep their native positions and that the unfolded regions do not fold into another, non-native globule.

The free energy of structure S_ν, relatively to that of the random coil, is taken as

$$F(S_\nu) = \varepsilon \sum_{(i<j+1)\in glob.S_\nu} \sum \delta_{ij} - T \cdot [(n_\nu - N)\sigma_1 + \sum_{loops\in S_\nu} S_{loop}]. \qquad (2)$$

The first sum is taken over all residues i, j keeping their native positions in S_ν. The last sum is taken over all the closed loops protruding from the native-like part of S_ν. δ_{ij} is the number of native atom-atom contacts (at a distance <5Å) between residues i and j (except i to i+1 contacts: they exist also in the coil); ε is the energy of one contact; n_ν is the number of unfolded residues; T is the temperature; σ_1 is the entropy difference between the coil and the native state of a residue (we take $\sigma_1 = 2.3R$ [16], R being the gas constant). At the point of equilibrium $F(S_0)=F(S_U)$, i.e., ε and T are

connected by equation $\varepsilon = -TN\sigma_1 / \sum\sum_{(i<j+1)\in S_0} \delta_{ij}$. The entropy spent to close a disordered loop between the still fixed residues *k* and *l* is estimated [17] as

$$S_{loop} = -\,{}^5/_2 R \ln|k - l| - {}^3/_2 R\,(r^2_{kl} - a^2) / (2Aa|k - l|)\,; \qquad (3)$$

here r_{kl} is the distance between the C_α atoms of residues *k* and *l*, a = 3.8Å is the distance between the neighbor C_α atoms in the chain, and *A* is the persistence length for a polypeptide (according to [20], we take *A* = 20Å).

To facilitate the computations, we sometimes have to use "chain links" of a few residues. These "links" somewhat limit the accuracy of our calculations, but not in a crucial way: they are still much smaller than the expected size of a folding nucleus in a vicinity of mid-transition between the folded and the unfolded phases (where the nucleus should include roughly 1/3 of the protein globule [17]).

2.3 Transition states on the protein unfolding pathways

Let us consider some unfolding pathway $w = (S_0 \to S_1 \to \ldots \to S_U)$; then $F^{\#}_w = \max\{F(S_0),F(S_1),\ldots,F(S_U)\}$ is the free energy of the TS ("free-energy barrier") at the pathway *w*. The most efficient kinetic pathway has the minimal (over all the pathways) TS free energy, $F^{\#}_{min} = \min_{possible\ w}\{F^{\#}_w\}$: this pathway passes from S_0 (the native state) to S_U (the coil) via the lowest free energy barrier. Let $S_{v-1} \in \{S_{v-1} \to S_v\}$ mean that S_{v-1} can be transformed into S_v in an elementary step (i.e., by removal of one link from the globular part of S_{v-1}). Thus,

$$F^{\#}_{min} = \min_{\substack{S_1,\ \ldots,\ S_{U-1} \\ S_1 \in \{S_1 \to S_2\} \\ \cdots\cdots \\ S_{U-2} \in \{S_{U-2} \to S_{U-1}\}}} \{\max\{F(S_0), F(S_1), \ldots, F(S_U)\}\}. \qquad (4)$$

Despite the huge number of possible pathways, $F^{\#}_{min}$ can be calculated either by a branch and bound (BB) technique [21], or by dynamic programming (DP) [22].

In application to our problem the key idea of the BB method is that having some estimate F_{limit} of the upper limit of the TS free energy, we can refrain from considering all the pathways from any state *S* where $F(S) \geq F_{limit}$. If some pathway *w* has been passed up to the end (from S_0 to S_U), we fix the F_{limit} to be equal to the free energy maximum $F^{\#}_w$ at this pathway. Then we follow this pathway back, up to that $S^{\#}_w$ structure which has $F^{\#}_w(S^{\#}_w) = F^{\#}_w$, make one more step back, and begin to explore the other pathway branches going from the "pre-maximum" state $S^{\#}_{w-1}$. If we can pass the new pathway *w** up to the end without violation of the limit F_{limit}, this means that we have found a new, lower barrier $F^{\#}_{w*}$ and the structure $S^{\#}_{w*}$

corresponding to the new TS. Then we make one step back from this structure and start to consider the other pathways from this state $S^{\#}_{w^*-1}$ forward (to S_U), and so on. If the movement along new pathway w^* brings us to a state S_{w^*} such that $F(S_{w^*}) \geq F_{limit}$, we make one step back from the S_{w^*} along the pathway w^* and start to consider other branches from this state S_{w^*-1}, and so on. In both the above cases we make one step back from some structure S with $F(S) \geq F_{limit}$. Moreover, we make one step back when we find a "completely explored" structure, i.e., such a structure that each pathway from it is either estimated (in the sense of maximal free energy at this pathway) or discarded, since the free energy at the pathway is above F_{limit}. The process ends when the initial structure S_0 becomes "completely explored". This algorithm guarantees that the found free energy barrier is the lowest of all the possible ones at the pathways leading from S_0 to S_U.

The same BB method can explore also the "suboptimal" pathways. To this end we take the F_{limit} already found for the optimal pathway, fix the free energy limit for suboptimal pathways equal to $F_{limit}+\Delta F$, and then consider the pathways not exceeding this limit using the above described method.

DP method also can solve equation like eq.(4). The algorithm is as follows. Let $b(S)$ be the altitude of the lowest free energy barrier at the pathways leading from S_0 to S inclusively (thus, $F^{\#}_{min} = b(S_U)$), and $q(S)$ be that at the pathways from S (exclusively) to S_U. The $b(S)$, $q(S)$ values are computed recursively:

$$
\begin{aligned}
& b(S_1) = \max\{F(S_0), F(S_1)\} \quad \text{for all intermediates } S_1; \\
& b(S_2) = \min_{S_1 \in \{S_1 \to S_2\}} \{ \max\{b(S_1), F(S_2)\} \} \quad \text{for all } S_2; \\
& \cdots\cdots \\
& F^{\#}_{min} = b(S_U) = \min_{S_{U-1}} \{ \max\{b(S_{U-1}), F(S_U)\} \}; \qquad (5)
\end{aligned}
$$

and

$$
\begin{aligned}
& q(S_{U-1}) = F(S_U) \quad \text{for all } S_{U-1}; \\
& q(S_{U-2}) = \min_{S_{U-1} \in \{S_{U-2} \to S_{U-1}\}} \{ \max\{F(S_{U-1}), q(S_{U-1})\} \} \quad \text{for all } S_{U-2}; \\
& \cdots\cdots \\
& q(S_1) = \min_{S_2 \in \{S_1 \to S_2\}} \{ \max\{F(S_2), q(S_2)\} \} \quad \text{for all } S_1; \qquad (6)
\end{aligned}
$$

($S_\nu \in \{S_{\nu-1} \to S_\nu\}$ means that S_ν is obtained from $S_{\nu-1}$ in one elementary step). Then

$$F^{\#}(S) = \max\{b(S), q(S)\} \qquad (7)$$

is the altitude of the lowest free energy barrier at the pathways leading from S_0 to S_U via each intermediate S.

The intermediate(s) with $F^{\#}(S)=F^{\#}_{min}$ give the transition state(s) with the minimal free energy. The intermediates with $F^{\#}(S)=F(S)$ give the ensemble $\{S^{\#}\}$ of all the possible passes over the free energy barrier dividing S_0 from S_U. $\{S^{\#}\}$ gives the utmost estimate of the variety of TS (it can be redundant since a pathway to the TS high in free energy can pass via some TS of the lower free energy). To outline the nucleus, we investigated both the single TS of the lowest free energy and the ensembles $\{S^{\#}\}$ of all possible transition states. In the last case we compute (for each residue i) the average fraction of the side chain native contacts preserved in the transition state ensemble $\{S^{\#}\}$ (we pay attention to the side chain contacts since just they are changed by mutations aimed to outline the TS experimentally):

$$\Phi(i) = \Sigma_{S\#} P(S^{\#}) [C(S^{\#},i) / C(S_0, i)] . \quad (8)$$

The sum is taken over ensemble $\{S^{\#}\}$; $P(S^{\#})=\exp(-F^{\#}(S^{\#})/RT)/[\Sigma_{S\#} \exp(-F^{\#}(S^{\#})/RT)]$ is the Boltzmann probability of $S^{\#}$ in the ensemble $\{S^{\#}\}$; $C(S^{\#},i)$ is the number of contacts between the side chain atoms of residue i keeping its native position and all atoms of the other residues having the native arrangement in state $S^{\#}$ (except those with next-neighbors of i: they exist in the coil as well); $C(S_0,i)$ is the number of these same contacts in the native structure. These Φ values have the same sense as the Φ_f values derived from protein engineering experiments ($\Phi_f(i)=1$ when the mutation of residues i affects the folding rate by changing the TS stability as strongly as that of the native protein, and $\Phi_f(i)=0$ when the mutation does not affect the folding rate [5]). The computed Φ and the experimental Φ_f values are compared to see the correlation of the theory with experiment.

To use DP in searching for the TSs at a network of folding-unfolding pathways, we have to restrict this network by $\sim 10^7$ intermediates. Therefore we use "chain links" consisting of a few residues: of two residues for proteins with less than 100 residues, and of four residues for larger proteins. To the same end we consider only the intermediates with no more than two closed loops in the middle of the chain plus the N- and the C-terminal disordered tails. These four unfolded regions should be enough to describe the unfolding of a protein up to $L \approx 100$ residues, since the estimated [17] number of coil regions in the folding nucleus is close to $L^{2/3}/6$.

2.4 Kinetic equations at the network of folding/unfolding pathways

The number n_i of protein molecules having state i (i = 0, 1 ,…, M, M+1 where "0" is the native state S_0, "M+1" the coil S_U, and "1",…, "M" the intermediates, see Fig.1) changes with time t according to usual kinetic equations

$$\frac{dn_i}{dt} = -\sum_{j=0}^{M+1} k_{ij} n_i + \sum_{j=0}^{M+1} k_{ji} n_j \quad (9)$$

where k_{ij} is the rate of transition from the i to the j state. These equations can be

solved in a quasi-stationary approximation [23], i.e., under the assumption that $dn_i/dt = 0$ for all the intermediates "1",…, "M", which is correct when all the intermediates have a high free energy and therefore low statistical weight as compared with the initial and the final states; thus, it is valid for "all-or-none" protein folding and unfolding. When all k_{ij} are given, the resulting rate $K_{0,M+1}$ of the "0"→"M+1" transition and $K_{M+1,0}$ of the "M+1"→"0" transition can be computed from solution of M linear equations. Then we can "mutate" a residue (by infinitesimal modification of its interaction energies with other residues), compute the modified $F_0 - F_{M+1}$, $K_{0,M+1}$ and $K_{M+1,0}$ values, and from them calculate (cf. Ref.5) the Φ value for this residue .

For transition rates we use an approximation usual for the Metropolis scheme of kinetic simulations based on Monte-Carlo method [4,24]:

$$k_{ij} = k_0 \times \begin{cases} 0, & \text{if transition } i \to j \text{ is physically impossible} \\ 1, & \text{if transition } i \to j \text{ is possible and } F_i \geq F_j \\ \exp[-(F_j - F_i)/RT], & \text{if transition } i \to j \text{ is possible and } F_i \leq F_j \end{cases}$$

(transition is "possible" when j is obtained from i by adding of one link to the native-like part of i, or by removing of one link from this native-like part). k_0 is the rate constant for a downhill (in free energy) step.

To have a limited size of the network of intermediates (and thus of equations) we use "chain links" of 4-8 residues, but do not restrict the number of loops.

2.5 Analyzed proteins and experimental data

All the calculations of folding nuclei in this paper are held for five proteins where the experimental Φ values have been obtained for many chain residues: barnase [25], chymotrypsin inhibitor 2 (CI2) [11,26], signal-transduction protein CheY [27], SH3 domain of src tyrosin-kinase transforming protein (src-SH3) [28], and SH3 domain of α-spectrin (α-SH3) [29]. Their 3D coordinates are taken from the PDB [19], files 1rnb.ent, 2ci2.ent, 3chy.ent, 1srm.ent and 1shg.ent, respectively.

The maximal number of contacts for each sort of amino acid is computed from a large set of 250 small (≤200 residues) non- or weakly homologous proteins [30] enriched with the above mentioned five proteins.

Experimental Φ_f values are taken from [11,25-29]. Since the rarely observed $\Phi_f < 0$ и $\Phi_f > 1$ values have no structural interpretation [26], and the errors in Φ_f values are about ±0.1, the rare values $\Phi_f < -0.1$ and $\Phi_f > +1.1$ are excluded. Φ_f is taken as 0 when $0 < \Phi_f < -0.1$, and as 1 when $1 < \Phi_f < 1.1$. When several Φ_f values are given for a residue, we average them. For barnase we take Φ_f as $1-\Phi_u$ [25] since its folding in pure water goes via a metastable intermediate, while its unfolding (u) and folding at the moderate denaturant concentrations is a two-state process studied here.

3 Results and Discussion

Before analysis of folding/unfolding pathways, it is worthwhile to elucidate the correlation of involvement of a residue into the folding nuclei with its role in the 3D protein structure.

Let θ_α be 1 if the residue is in an α-helix and 0 if not, θ_β be 1 if the residue is in a β-sheet and 0 if not, and C^*_d (eq.(1)) be a fraction of the residue's atom-atom contacts (up to distance d) from that maximally possible for given amino acid.

$C^*_{5Å}$ characterizes the involvement of a residue into the hydrophobic core of the protein. $C^*_{5Å}$ correlates with the experimental Φ_f value at the average level of only 17% (Table 1). The absolute number of residue contacts $C_{5Å}$ correlates with the Φ_f value even worse. A weak correlation of the number of native contacts with the Φ_f values has been mentioned in [10] and further investigated in [31].

$\theta_\alpha+\theta_\beta$ characterizes the involvement of a residue into the secondary structure. It correlates with the Φ_f only a little better, at the average level of 19% (Table 1).

The generalized value $A=\theta_\alpha+\lambda\theta_\beta+\mu C^*_d$, where λ, μ and d are the optimized parameters, correlates with the Φ_f value at the average level of 26% only (Table 1), – although the values d=9.2Å, λ=0.5 and μ=4.1 have been optimized (to increase correlation of A with Φ_f) at the same five proteins (which must overestimate the resulting correlation).

All this suggests that it is impossible to predict participation of residues in the folding nucleus by just the participation of residues in the protein hydrophobic core or the secondary structure without modeling of its folding/unfolding.

Table 1. Coefficient of correlation between experimental Φ_f values and involvement of residues in the hydrophobic core and in the secondary structure of proteins

Protein	CI2	Barnase	CheY	src-SH3	α-SH3	Average
Number of experimental points	39	29	27	15	6	-
Involvement in hydrophobic core ($C^*_{5Å}$)	0.16	0.01	0.26	0.05	0.83	0.17
Involvement in secondary structure ($\theta_\alpha+\theta_\beta$)	0.41	0.17	0.15	-0.20	0.06	0.19
Involvement in hydrophobic core and secondary structure ($\theta_\alpha+0.5\theta_\beta+4.1C^*_{9.2Å}$)	0.36	0.08	0.42	-0.18	0.87	0.26

3.1 Results of investigation of the folding/unfolding pathways.

Table 2 shows that, on the average, the correlation of the "optimal" (according to calculation) folding nuclei with the experimental Φ_f values is about 30%, i.e., it is somewhat higher than that of Φ_f values with the residue involvement in the hydrophobic core and in the secondary structure. Although the correlation itself virtually does not depend on the size of links used in computations, the exact position of the "optimal nucleus" in the chain is rather sensitive to all the computational details, including the link size (Fig.2). This shows that one should not

consider only one, even the "optimal" TS: the "suboptimal" transition states are very close to the "optimal" state (Fig.2), their free energies exceeding that of the "optimal" folding nucleus only by a small fraction of *RT*. In other words, the choice of the sole "optimal" nucleus depends on all details of calculations.

Table 2. Computed folding nuclei and their correlation with the experimental Φ_f values

Protein	CI2	Barnase	CheY	src-SH3	α-SH3	Average
Number of experimental points	39	29	27	15	6	-
Number of amino acid residues	64	109	128	56	57	82,8
Optimal TS, $\Delta F^{\#}$ (in *RT* units) [BB method*]	13.5	19.4	10.7	14.1	13.1	14.1
Optimal TS [by BB*], correlation with Φ_f	0.37	0.16	0.56	-0.24	0.50	0.29
Optimal TS [by DP**], correlation with Φ_f	0.35	0.16	0.52	-0.12	0.50	0.29
Ensemble of all the TSs [by DP**], correlation with experimental Φ_f	0.48	0.49	0.60	-0.01	0.54	0.45

* "Chain link" consists of 2 residues for barnase (where computations with 1-residue link overflows the computer memory), and of 1 residue for other proteins;
** "Chain link" consists of 4 residues for barnase and CheY, and of 2 residue for the other proteins.

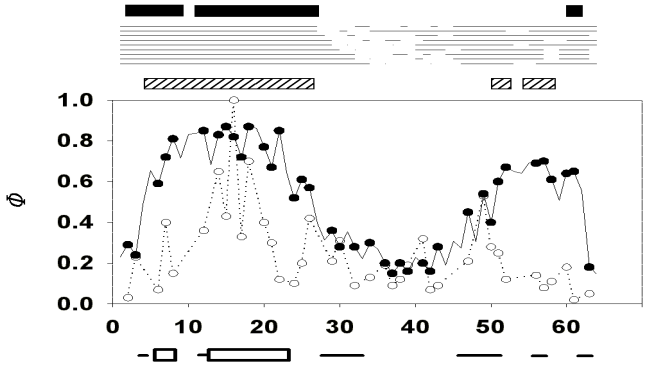

Fig.2. At the top: location of the "optimal" folding nucleus (black rectangles) and of some "suboptimal" nuclei (thin lines) in the CI2 chain. The computation was done by branch and bound method (where a "link" consisted of one residue). Free energies of the suboptimal nuclei, about 13.4817 *RT*, are only a little higher than that of the optimal nuclei, 13.4816 *RT*. The hatched rectangles (below) show the location of the optimal folding nucleus computed with two residues in a link. The plot (solid line with filled circles where a comparison with experiment is possible) shows the theoretical Φ values calculated for the transition state ensemble in CI2 using dynamic programming (with two residues in a link); for more plots of this kind, see Galzitskaya & Finkelstein, *Proc. Natl. Acad. Sci USA*, in press. The experimental Φ_f values are shown with open circles (connected with dotted line for better presentation). The rectangles and lines below the plot show the location of the native α-helices and the β-strands in the protein chain.

Thus, it is better (Table 2 and Fig.2) to consider the ensemble $\{S^{\#}\}$ of all the folding/unfolding transition states. This ensemble is more appropriate to compute by DP than by the BB technique. However, the application of DP requires a limitation

of the considered set of intermediates. The volume of this set depends on the limitations used in computations. The possible limitations can be estimated using a less demanding BB method. The "optimal" folding nuclei found in this way in small proteins usually include only one-two, very rarely three, closed loops protruding from the nucleus. This shows the limitation of the loop number (≤2) that can be allowed in our calculations of the Φ values by dynamic programming.

In solving kinetic equations for a network of folding pathways we get the Φ values close to those obtained by DP for a whole TS ensemble (cf. Fig.3 and Fig.2). Like experiment, both theories show high Φ values for the N- and C-terminal parts of the CI2 chain (this means that they are involved in the nucleus), but the peaks for the theoretical Φ's are broader than those for the experimental Φ_f's. This is probably due to the neglected specificity of atomic contacts and to the rough estimate of loop entropy in our calculations.

Generally, the computed Φ values are not very sensitive to such details as small changes of contact energies or a modification of the link size (usually, the results are essentially the same for 2-, 3- or 4-residue links). The only exception is src-SH3. It has the worst predicted Φ values when the DP computations are held with 2-residue links (*Corr* = -1%, the only example of negative correlation in our practice, see Table 2), but *Corr* = 54% when 3-residue links are used.

All the considered proteins have the two-state transitions between the native globule and the random coil in a vicinity of the denaturation point (a metastable folding intermediate is observed for some of them only when the native state is very stable). The developed theories refer just to these two-state transitions. It seems that they can be applied also to the molten (or swollen) globule-native state transitions for the cost of some modification of the energy and entropy terms in equations (2), (3). However, they cannot be applied to the coil-molten globule transition when the 3D structure of the molten globule is not known.

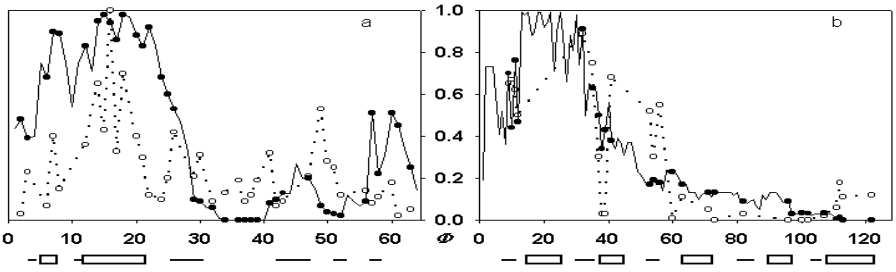

Fig.3. Φ values calculated from kinetic equations for the network of CI2 (a) and CheY (b) folding/unfolding pathways (with 4 residues in a "chain link" in CI2, and with 8 in CheY). For CI2, the computed folding rate $K=1.2\times10^{-7}\,k_0$, i.e., the free energy barrier is $16RT$. For CheY, $K=9\times10^{-7}\,k_0$, and the barrier is of $14RT$. With $k_0 \sim 10^7$ sec^{-1} (a reasonable rate estimate for folding/unfolding of a "link" of a few residues [35]), K is of the order of ~1 sec^{-1} for both proteins, in a reasonable concordance with their folding/unfolding rates (at the considered point of thermodynamic equilibrium of the native

structure and the coil) [26,27]. The correlation between the theoretical (-•-) and the experimental (o) Φ values is 48% for CI2 and 78% for CheY.

4 Conclusions

This study shows that the coarse-grained model of sequential protein folding [17] gives a possibility to outline the folding nucleus. And although the nucleus is outlined by this model more crudely than by MD simulations reported in [10,11] for CI2, the presented approaches have two important advantages over the MD simulations: they do not need neither additional experimental data nor additional speculations to single out the TS, and they are fast. Our next aim is to do them more precise using more precise estimates of inter-residue interactions.

An overview of the calculated TS ensembles shows that many of the parallel transition states, though of nearly equal free energy, have substantial variations in size and positions. This result correlates with the suggestion that a 3D structure can fold using various folding nuclei [29,32]. It should be also mentioned that the found globular parts of the TS structures are not very small: usually, they include from one third to a half, even up to two thirds of all the chain residues. Theoretically [17], such large and not too specific nuclei must be typical of folding (and unfolding) close to the point of thermodynamic equilibrium between the native globule and the coil. However, a stabilization of the native structure must make the nuclei smaller and more specific. This is suggested also by comparison of the results of folding simulations [32] and [4,6,7] held under different conditions, as well as by some experimental results [33,34].

In the examined proteins the semi-folded structures have high free energies; this is consistent with the two-state all-or-none transition between the native and the unfolded state. Most of these semi-folded structures have a very high free energy of many tens or even hundreds of *RT* units. However, the calculation finds the passages through this high free energy landscape where the free energy of the maximum exceeds that of the native (or coil) states by only 10 - 20 *RT* (Table 2). This is consistent with the estimates [17] for proteins of the examined size (of 60-120 residues; however, it is noteworthy that Table 2 shows no direct dependence of the TS free energy on the protein size). Such relatively low free energy barriers allow these proteins to fold within milliseconds or seconds [17] in a reasonable semi-quantitative concordance with experiment [25-29]. The computed transition states look compact and contain small number of protruding loops.

Acknowledgments

We are grateful to the Howard Hughes Medical Institute International (Research Scholar Award No. 75195-544702) and to the Russian Basic Research Foundation (Award No. 98-04-49303) for support of this work.

References

1. A.R. Fersht, *Curr. Opin. Struct. Biol*. **5**, 79 (1995)
2. A.R. Fersht, *Curr. Opin. Struct. Biol*. **7**, 3 (1997)
3. C.M. Dobson and M. Karplus, *Curr. Opin. Struct. Biol*. **9**, 92 (1999)
4. V.I. Abkevich et al, *Biochemistry* **33**, 10026 (1994).
5. A. Matouscheck et al, *Nature* **346**, 440 (1990)
6. E. Shakhnovich, V. Abkevich and O. Ptitsyn, *Nature* **379**, 96 (1996)
7. L.A. Mirny et al, *Proc. Natl. Acad. Sci USA* **95**, 4976 (1998)
8. J.J. Portman, S. Takada and P.G. Wolynes, *Phys. Rev. Lett.* **81**, 5237 **(**1998)
9. M.C. Demirel et al, *Protein Science* **7**, 2522 (1998)
10. A. Li and V. Daggett, *J. Mol. Biol*. **257**, 412 (1996)
11. V. Daggett et al, *J. Mol. Biol*. **257**, 430 (1996)
12. A. Caflisch and M. Karplus M. *J. Mol. Biol*. **252**, 672 (1995)
13. C.L. Brooks III et al, *Proc. Natl. Acad. Sci. USA* **95**, 11037 (1998)
14. A.V. Finkelstein, *Protein Eng*. **10**, 843 (1997)
15. P.T. Landsberg, ed., *Problems in Thermodynamics and Statistical Physics* (PION, London, 1971).
16. P.L. Privalov, *Adv. Protein Chem.* **33**, 167 (1979)
17. A.V. Finkelstein and A.Ya. Badretdinov, *Fold. Des*. **2**, 115 (1997)
18. W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983)
19. F.C. Bernstein et al, *Eur. J. Biochem*. **80**, 319 (1977)
20. P.J. Flory, *Statistical Mechanics of Chain Molecules.* (Interscience, New York, 1969)
21. V. Lipskii, *Combinatorics for programmers* (Mir, Moscow, 1988).
22. A. Aho, J. Hopcroft and J. Ullman, *The Design and Analysis of Computer Algorithms* (Addison-Wesley, Reading, MA, 1976)
23. N.M. Emmanuel and D.G. Knorre. *Lectures in chemical kinetics.* (Vysshaya Shkola, Moscow, 1984)
24. H.J. Hilhorst and J.M. Deutch, *J. Chem. Phys.* **63**, 5153-5161 (1975)
25. L. Serrano et al, *J. Mol. Biol*. **224**, 805 (1992)
26. L.S. Itzhaki et al, *J. Mol. Biol*. **254**, 260 (1995)
27. E. Lopez-Hernandez and L. Serrano, *Fold. Des.* **1**, 43 (1996)
28. V.P. Grantcharova et al, *Nat. Struct. Biol*. **5**, 714 (1998)
29. A.R. Viguera et al, *Nat. Struct. Biol*. **3**, 874 (1996)
30. C.-T. Zhang et al, *Protein Eng*. **11**, 971 (1998)
31. B. Nolting, *J. Theor. Biol.* **197**:113 (1999)
32. D.K. Klimov and D. Thirumalai, *J. Mol. Biol*. **282**, 471 (1998)
33. M. Oliveberg, *Acc. Chem. Res*. **31**, 765 (1998)
34. S.E. Jackson, *Fold. Des.* **3**, R81-R91 (1998)
35. S. Williams et al., *Biochemistry* **35**, 691 (1996)