# Understanding the Folding Rates and Folding Nuclei of Globular Proteins

Alexei V. Finkelstein*, Dmitry N. Ivankov, Sergiy O. Garbuzynskiy and Oxana V. Galzitskaya

*Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia*

**Abstract:** The first part of this paper contains an overview of protein structures, their spontaneous formation ("folding"), and the thermodynamic and kinetic aspects of this phenomenon, as revealed by *in vitro* experiments. It is stressed that universal features of folding are observed near the point of thermodynamic equilibrium between the native and denatured states of the protein. Here the "two-state" ("denatured state" ↔ "native state") transition proceeds without accumulation of metastable intermediates, but includes only the unstable "transition state". This state, which is the most unstable in the folding pathway, and its structured core (a "nucleus") are distinguished by their essential influence on the folding/unfolding kinetics. In the second part of the paper, a theory of protein folding rates and related phenomena is presented. First, it is shown that the protein size determines the range of a protein's folding rates in the vicinity of the point of thermodynamic equilibrium between the native and denatured states of the protein. Then, we present methods for calculating folding and unfolding rates of globular proteins from their sizes, stabilities and either 3D structures or amino acid sequences. Finally, we show that the same theory outlines the location of the protein folding nucleus (i.e., the structured part of the transition state) in reasonable agreement with experimental data.

## INTRODUCTION

The aim of this paper is to outline the modern understanding of the physical principles of protein structure self-organization.

A protein is a heteropolymer built of amino acid residues. It has a chemically regular backbone chain and a unique (for each protein) sequence of 20 kinds of side groups; sometimes, it also includes "cofactors", which are usually small molecules, and chemical modification of some amino acids [1].

Before considering protein physics, it is useful to remind a reader that proteins exist under various environmental conditions which leave an obvious mark on their structures. Roughly, according to the "environmental conditions" and general structure, proteins can be divided into three large groups [2].

Fibrous proteins form vast aggregates; their structure is usually maintained mainly by interactions between different polypeptide chains.

Membrane proteins exist in water-lacking membranes. Their intramembrane portions have highly regular three-dimensional (3D) structure (like fibrous proteins) but are restricted in size by the membrane thickness (30–40Å).

Water-soluble globular proteins that live in aqueous environments are the most numerous and well studied.

In this review we will deal only with globular proteins. Moreover, we will concentrate mostly on relatively small, "single-domain" proteins that form one compact protein globule. A "single-domain" structure is typical of small, water-soluble globular proteins. Large proteins usually consist of two, three or even more domains [2,3].

*Address correspondence to this author at the Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia; Tel./Fax: +7-495-632-7871; E-mail: afinkel@vega.protres.ru

The chain of an "operating" protein is usually folded into a "*native*" 3D structure, which is strictly specified except for small fluctuations and (sometimes) small rearrangements, and which exists under normal biological conditions but decays under the action of various denaturants, such as temperature, acid, some chemicals like urea, etc. Some (~10%) protein chains, however, have no fixed structure by themselves, but obtain it by interacting with other molecules [4].

Protein physics is grounded on three fundamental experimental facts: (i) many proteins have well defined (except for small fluctuations) three-dimensional structures [5-7]; (ii) many protein chains are capable of self-organization, i.e., they form their native structures spontaneously in an appropriate environment [8,9], and (iii) the native state of many proteins is separated from the unfolded state of the chain by an "all-or-none" phase transition [10]. The last point ensures the robustness of protein action: as a bulb, the protein either has a correct structure and works correctly, or does not work at all.

## 1. PHASE TRANSITIONS IN PROTEIN MOLECULES

### 1.1. Reversible Denaturation of Protein Structures

Under biological conditions the native structure of a protein is rather "solid". However, depending on ambient conditions, the most stable state of a protein molecule may be not solid but molten or even extremely swollen, "unfolded": then the protein "denatures" and loses its native, "working" 3D structure.

Denaturation of a protein is usually reversible: many protein chains are capable of self-organization (usually called "renaturation" or "folding"), i.e., they form their native structures *spontaneously* in an appropriate environment [8,9]. This spontaneous self-organization not only happens to natural proteins produced by a living organism [8], but also to proteins whose amino acid sequences (copied from those of

1389-2037/07 $50.00+.00 

natural proteins) have been synthesized chemically in a test tube [9].

Folding as a phenomenon has been known since Anfinsen's famous experiments in the 1960's [8]. Since then, spontaneous folding has been observed for hundreds of proteins [11]. It is now known that a protein can renature (when the ambient conditions return to "physiological" ones) if it is not too large and has not been subjected to substantial chemical modifications after the initial *in vivo* folding (and if the protein solution is sufficiently diluted to avoid aggregation). In this case, a "mild" (without chemical decay) destruction of the protein's native structure (by temperature, denaturant, etc.) is reversible, and the native structure is spontaneously restored after environmental conditions have become normal.

Usually protein denaturation and renaturation are studied *in vitro* (in a test tube); then denaturation is caused by an abnormal, "non-physiological" temperature or by an excess of a denaturant (like $H^+$, $OH^-$ or urea). However, decay of the "solid" protein structure (and its subsequent refolding) can also occur in a living cell, e.g., during trans-membrane transport of proteins.

The spontaneous folding, i.e., the reversibility of protein denaturation, shows that the entire information necessary to build up the protein 3D structure is contained in its amino acid sequence, and that the protein structure itself (to be more exact, the structure of a protein that is not too modified and not too large) is thermodynamically stable. This allows one to use thermodynamics to study and describe de- and renaturation transitions.

Denaturation and folding of water-soluble globular proteins is the most studied, and we will speak about them only.

It is well established that denaturation of small native proteins is a cooperative transition with a simultaneous, abrupt change of various characteristics of the molecule. Narrow transition regions suggest that the transition embraces many amino acid residues.

Moreover, denaturation of a single-domain protein occurs as an "all-or-none" transition [10,12]. The latter means that the transition embraces the domain as a whole, and that only the initial (native) and the final (denatured) states amount to visible quantities, while "semi-denatured" states are unstable and practically absent. [Though, of course, they do exist to a very small extent, since a native molecule cannot come to its denatured state without passing the intermediate forms, and their presence has a crucial effect on kinetics of the transition, which we will discuss below.]

The "all-or-none" transition is a microscopic analog of the first-order phase transitions in macroscopic systems (e.g., crystal melting). However, unlike the true phase transitions, the "all-or-none" transitions in proteins have a non-zero temperature width, since this transition embraces a microscopic system. It should be specified that the "all-or-none" denaturation actually concerns small proteins and separate domains of large proteins, while denaturation of a large protein is a sum of denaturations of its domains [13,14].

To prove that melting is an "all-or-none" transition, one has to compare (1) "effective latent heat" of transition calculated from its width (i.e., the amount of heat consumed by one independent melting unit) with (2) "calorimetric heat" of this transition, i.e., the amount of heat consumed by one melting protein molecule [10]. Denaturation of proteins is usually accompanied by a large heat effect, ~1 kcal/mol of amino acid residues; however, the native state of a protein chain is more stable than its unfolded state by no more than a few kcal/mol even under physiological conditions, where the native state is the most stable.

### 1.2. What Do Denatured Proteins Look Like?

Numerous *thermodynamic* experiments have shown that there are no cooperative transitions within the denatured state of a protein molecule. Therefore, it was initially assumed that the denatured protein is always a very loose random coil (as it is in a "very good" solvent like concentrated solution of urea [11]).

However, *structural* studies of denatured proteins reported on some large-scale rearrangements within the denatured state, that is, on some stable "intermediates" between the completely unfolded coil and the native state of proteins (Fig. **1**).

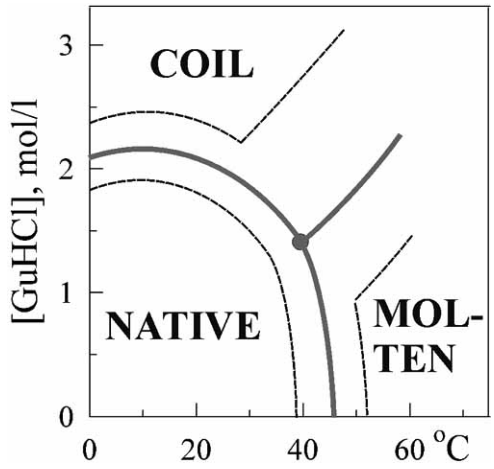


**Fig. (1).** Phase diagram of conformational states (at pH 1.7) of a single-domain protein lysozyme at various temperatures in solution of guanidine hydrochloride (GuHCl) denaturant of various concentrations: the solid NATIVE state, the completely unfolded COIL, and a more compact temperature-denatured state (MOLTEN). The solid line corresponds to the mid-transition, the dashed lines outline the transition zones (from the proportion ≈9:1 in favor of one state to ≈1:9 in favor of another). One can see that the COIL–MOLTEN transition is much wider than the others. Adapted from [15]. Such phase diagram is common for other denaturants and proteins.

Apart from biochemical activity, only two properties always abruptly change during denaturation. These are: (1) the ordering of the environment of aromatic side chains observed by near-UV CD and by NMR, and (2) the rigidity of the globular structure followed (using NMR) by exchange of hydrogens (H) of the protein's polar groups for deuteriums (D) of water, and by acceleration of the protein chain proteolysis [16-18].

A protein chain looks like a random coil in a "very good" solvent, such as a concentrated solution of urea or GuHCl; then, its hydrodynamic volume is proportional to the chain

length to the power 3/2 or close to 3/2. However, in poor solvents (e.g., water) the volume of the denatured globule is often only a little larger than the volume of the native protein. Thus, the above studies have revealed an intermediate of protein unfolding and folding, which is now known as the "*molten globule*" [18] (Fig. **2**).

For very many (though not for all) proteins, the "molten globule" arises from a moderate denaturing impact upon the native protein, and decays (turns into a random coil) only under the impact of a concentrated denaturant. The molten globule-like state often occurs after temperature denaturation ("melting"), and this melting has been always observed to be the "all-or-none" transition. The molten globule usually does not undergo cooperative melting with increasing temperature (see Fig. **1**), but its unfolding caused by a strong denaturant looks like a cooperative S-shaped transition.

However, some proteins (especially small ones) unfold directly into a coil without the mediating molten globule state; and many other proteins are converted into the molten globule by some denaturing agents (e.g., by temperature or by acid), while other agents (e.g., urea) directly convert them into a coil. In all these cases, though, the decay of the native state is "all-or-none" transition between the native state and the molten globule (not coil!) state, i.e., via the first order phase transition between these states [18].

It has been shown that the "all-or-none" folding/unfolding phase transition requires a selected amino acid sequence that only provides a large energy gap between the native and all the other, structurally dissimilar conformations of the protein chain [22-24]. As a result of "all-or-none" transition, the protein tolerates, without a change, modification of ambient conditions up to a certain limit, and then melts altogether, like a solid body. This resistance and hardness of the protein, in turn, provides reliability of its biological function, and therefore must be maintained by biological evolution [2].

### 1.3. Protein Folding *In Vivo*

In a living cell, a protein is synthesized by a ribosome that makes a protein chain (whose sequence is encoded by mRNA) residue by residue, from its N- to C-end, and not quite uniformly: there are temporary rests of the synthesis at the "rare" codons (they correspond to tRNAs which are rare in the cell, and these codons are rare in the cell's mRNAs, too). It is assumed that the pauses may correspond to the boundaries of structural domains that can help a quiet maturation of the domain structures. The biosynthesis takes about a minute and yielding of a "ready" folded protein lasts just as long: the experiment does not see any difference [1,25].

Some enzymes, like prolyl-peptide- or disulfide-isomerases accelerate *in vivo* folding. They catalyze slow, if unaided, *trans*↔*cis* conversions of prolines and the formation (and decay) of chemical S-S bonds between cysteine amino acid residues.

Protein chains fold under the protection of special proteins, called chaperons. These are the cell's trouble-shooters that fight aggregation, since, in a cell, folding takes place in a highly crowded molecular environment. There is no reason to assume, though, that anything other than the amino acid sequence determines protein conformation in the cell [26,27].

It looks as though the biosynthetic machinery (ribosomes + chaperones + …), besides synthesizing the protein chain, serves only as a kind of incubator, which does not determine the protein structure (at least if the protein is not very large and does not consist of many domains) but rather provides "hothouse" conditions for its maturation, – just like a usual incubator helps a nestling to develop but does not determine what will be developed, a chicken or a duckling.

Unfortunately, it is difficult to follow the *in vivo* folding of a nascent protein chain against the background of the huge ribosome. It is known, though, that the first synthesized domains of multi-domain proteins are able to fold before the

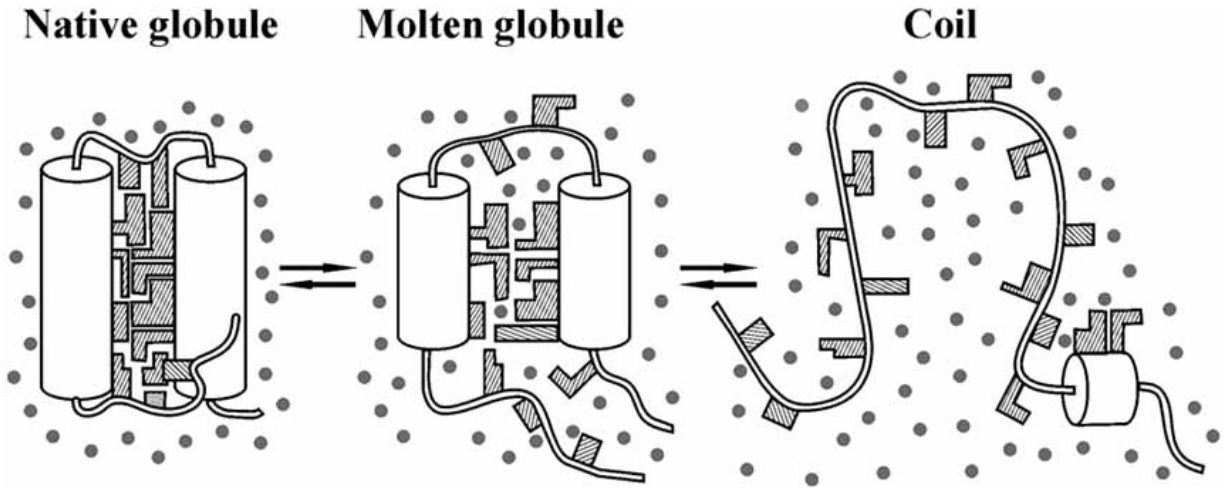


**Fig. (2).** Schematic model of the molten protein globule [18-21] in comparison to the native protein and the coil. For simplicity, the protein is shown to consist of only two helices and irregular loops. The backbone is covered with numerous side chains (dashed). Reinforced by H-bonds, the secondary structures are stable until the globule is "dissolved" by a solvent. Usually, water molecules are unable to do this without a strong denaturant. In the molten globule, side chains lose their close packing, but acquire freedom of movement, i.e., they lose energy but gain entropy. Water molecules (◦) move into pores of the molten globule (that appear when the close packing is lost), but, until the denaturant is not too strong, cause no further decay of the globule; a stronger denaturant converts the globule into the coil.

biosynthesis of the whole chain is completed [1,27]. However, there is virtually no data on *in vivo* folding events for single-domain proteins.

Therefore, most experiments on protein folding are done *in vitro*.

### 1.4. Protein Folding *In Vitro*

In about 1960, a remarkable discovery was made: it was shown that a globular protein is capable of spontaneous folding *in vitro* [8]. If a protein chain has not been heavily chemically modified after the initial (*in vivo*) folding, and is then gently (without chain damaging) unfolded by temperature, denaturant, etc., the protein spontaneously "renatures", i.e., restores its activity and structure after solvent "normalization". True, effective renaturation requires a careful selection of experimental conditions; otherwise, aggregation (including famous amyloid formation [28,29]) can prevent the protein from folding.

Furthermore, it was demonstrated [9] that the protein chain which had been synthesized chemically, without any cell or ribosome, and placed in the proper ambient conditions, folds into a biologically active protein.

The phenomenon of spontaneous folding of protein native structures allows us to detach, at least to a first approximation, the study of protein folding physics from the study of protein biosynthesis.

Protein folding *in vitro* is the simplest (and therefore, the most interesting for a physicist) case of pure *self*-organization: here nothing "biological" (except for the sequence!) helps the protein chain to fold.

#### *1.4.1. The Levinthal Paradox*

The ability of proteins (and RNA) to fold spontaneously immediately raised a fundamental problem that has come to be known as the Levinthal paradox [30]. It reads as follows.

On the one hand, the same native state is achieved by various folding processes: *in vivo* on the ribosome, *in vivo* after translocation through the membrane, *in vitro* after denaturation with various agents… The existence of the spontaneous and correct folding of chemically synthesized protein chains suggests that the native state is thermodynamically the most stable under "biological" conditions.

On the other hand, a chain has zillions of possible conformations (at least $2^{100}$ for a 100-residue chain, since at least two conformations are possible for each residue), and the protein can "feel" the right stable structure only if it is achieved exactly, since even a 1Å deviation can strongly increase the chain energy in the closely packed globule. Thus, the chain needs at least ~$2^{100}$ picoseconds, or ~$10^{10}$ years to sample all possible conformations in its search for the most stable fold.

Then, a question arises: how can the chain find its most stable structure within a "biological" time (minutes) at all?

The paradox is that, on the one hand, the achievement of the same (native) state by a variety of processes is (in physics) clear-cut evidence of its stability. On the other hand, Levinthal's estimate shows that the protein simply does not have enough time to prove that the native structure is the most stable among all possible structures!

In order to solve this paradox, Levinthal suggested the existence of specific folding pathways, and hypothesized that the native fold is simply an end of the protein-specific pathway rather than the most stable fold of its chain. Should this pathway be narrow, only a small part of the conformational space would be sampled, and the paradox would be avoided. In other words, Levinthal suggested that the native protein structure is under kinetic rather than under thermodynamic control, i.e., that it corresponds not to the global, but rather to the easily accessible, free energy minimum.

#### *1.4.2. Folding Pathways and Folding Intermediates*

The question as to whether the protein structure is under kinetic or thermodynamic control is not a purely speculative question. It is raised again and again when one faces practical problems of protein physics and engineering. For example: when trying to predict protein structure from a sequence, for what must we look? The most stable, or the most rapidly folding structure? When designing a de novo protein, what must we do? Should we maximize the stability of the desired fold or create a rapid pathway to this fold?

A discussion on protein folding mechanisms started immediately after discovery of spontaneous folding. It seems that the first proposed hypothesis was by Phillips, who suggested that the folding nucleus is formed by the N-end of the nascent protein chain, and that the remaining part of the chain wraps around it [31]. This appealing hypothesis is present in some works up to now. However, it has been refuted experimentally, as far as single-domain proteins are concerned. The elegant works by Goldenberg & Creighton have shown that the N-terminus has no special role in *in vitro* folding: it is possible to glue the ends of the chain of a small protein with a peptide bond, and nevertheless it folds into the correct 3D structure [32]. Moreover, it is possible to cut this circular chain to make a new N-end at the former middle of the chain; the protein folds, nevertheless, to the former native structure. Nowadays, protein engineering routinely produces circularly permuted proteins.

In an effort to solve the folding problem, Ptitsyn proposed a model of stepwise protein folding [33] (Fig. **3**). Later given the name "framework model," this hypothesis stimulated investigation of folding intermediates. It postulated a stepwise involvement of different interactions in the protein structure formation, and it stressed the importance of rapidly folded $\alpha$-helices and $\beta$-hairpins at the initial folding steps, gluing of these helices and hairpins into a native-like globule, and crystallization of the final structure within this globule at the last step of folding.

The cornerstone of this concept was the then hypothetical and now well known folding intermediate, the "molten globule", which was later discovered and studied first as the equilibrium state of a "weakly denatured" protein [16,32] and then as a kinetic folding intermediate [35].

Experiment shows that different properties of the native protein have two quite different rates of restoring. Nearly-native volume and secondary structure restore within a second, while side-chain order, tertiary structure and biochemi-

cal activity take minutes to restore. This is evidence for the accumulation of some "intermediate" state of the protein molecule (as shown, the molten globule [18]) at the beginning of the folding process.

The molten globule is an early folding intermediate in the *in vitro* folding of many proteins at physiological conditions [18,36,37]. It takes a few milliseconds to form, while the complete restoration of native properties of a 100 – 300 residue chain can take seconds for some proteins and hours for others. Thus, the rate-limiting folding step concerns formation of the native "solid" protein from the molten globule rather than formation of the molten globule from the coil.

The molten globule is not the only intermediate observed in protein folding. The "pre-molten" globule (that also fits the "framework model") was observed [38] to precede the molten globule formation. As a kinetic intermediate it was discovered with the use of ultra-fast (sub-millisecond) measuring techniques [39,40]. In addition, proteins with disulfide bonds allow for the trapping of various intermediates indicative of the order of S-S bond formation, etc [11].

The "kinetic control" hypothesis initiated very intensive studies of folding intermediates. Actually, it was clear almost from the very beginning that the metastable intermediates are not obligatory for folding (since the protein can fold also near the point of equilibrium between the native and denatured states, where the transition is of the "all-or-none" type, which excludes any metastable intermediates). The idea was, though, that the intermediates, if trapped, would help to trace the folding pathway, like intermediates in a complicated (bio)chemical reaction trace its pathway. This was, as it is now called, "chemical logic". However, this logic worked only in part when it came to the protein folding. The intermediates (like molten globules) were found for many proteins, but the main question as to how the protein chain can rapidly find its native structure among zillions of alternatives remained unanswered.

### *1.4.3. "Two-State" and "Multi-State" Protein Folding*

Progress in the understanding of protein folding [41,42] has been achieved just by investigation of those proteins, which fold without "unnecessary complications" (previously widely used to trace the folding pathway): without accumulation of any intermediates at the folding pathways, without cis-trans proline isomerization, and without S-S bond formation. The folding (and the unfolding) kinetics look very simple in this case: all the properties of the native (or denatured) protein are restored synchronically, following the single-exponential kinetics [43]. For some proteins, this simplicity is observed in a wide range of conditions, including the denaturant-free water ("biological zone" in Fig. **4**), the zone of the reversible thermodynamic transition between two phases (the native and the denatured state) and the unfolding zone; these proteins obtained a name of "two-state proteins". For the other, "multi-state" proteins, the two-state folding occurs only in the transition zone, if any, while the unfolding demonstrates a "two-state" manner (Fig. **4**). Usually, the complicated folding demonstrates three phases, and the corresponding proteins obtained a name of "three-state proteins" [41-45]. Thus, the most universal features of folding (and unfolding) can be observed just in and around the transition zone, while the moving of this zone towards the "biological" conditions reveals individualities of various proteins (which are the "unnecessary complications", when we try to understand the basics of protein folding).

The above statement looks, in a sense, paradoxical. Indeed, what can we get from investigation of folding (or unfolding) in the transition zone, where we cannot accumulate any transition intermediates? The answer is: just here we can most readily, though indirectly, observe the folding transition state, whose stability (or, more exactly, instability) determines the folding (and unfolding) rate [41-44,46-48]. The transition state corresponds to the free energy maximum on the folding/unfolding pathway, – or, it is better to say, to the free energy saddle point on the network of these pathways

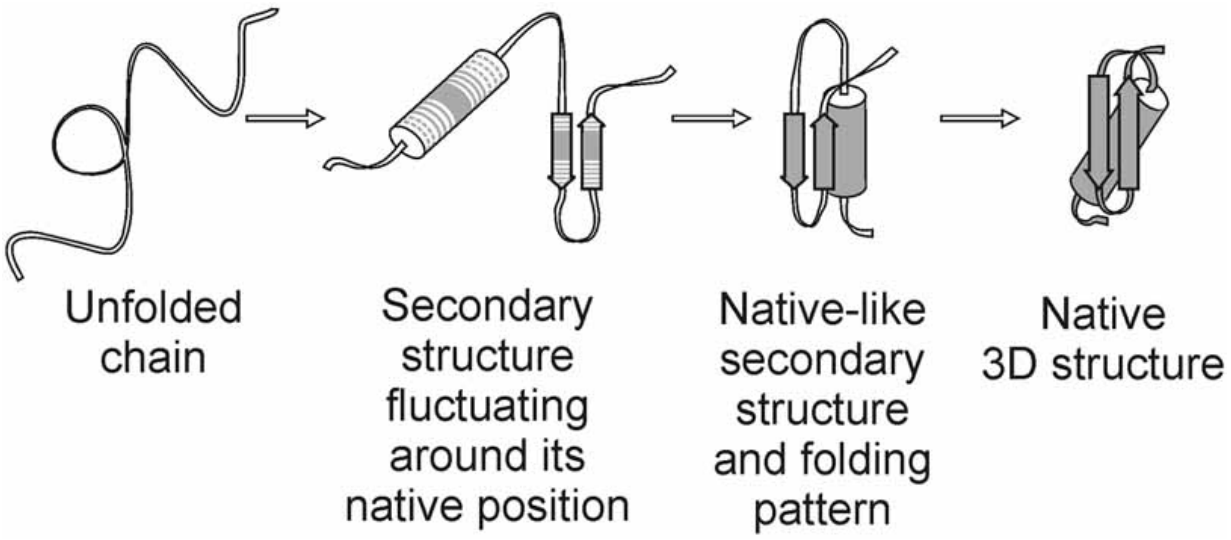


**Fig. (3).** Framework model of stepwise folding [33]. The secondary structures are shown as cylinders (α-helices) and arrows (β-strands). Both predicted intermediates have already been observed; the first is now known as the "pre-molten globule" and the other as the "molten globule" [18].

(see Fig. **10** below). The folded part of the transition state is called the "folding nucleus", and the folding pathway via formation of a nucleus (which usually consists of amino acid residues remote in protein chain [49,50]) obtained a name of "nucleation-condensation" mechanism of folding.

### *1.4.4. Folding nucleus.*

The "folding nucleus" plays a key role in protein folding: its instability determines the folding and unfolding rate-limiting steps. It should be stressed that the folding nucleus is **not** the molten globule, although some of their characteristics may be similar [48]: the nucleus corresponds to the free energy maximum, while the molten globule which is observed as a folding intermediate corresponds to a local free energy minimum [18]. It has been shown that the nucleus looks like part of the 3D structure of the native protein [44,48].

So far, there is only one, very difficult experimental method to identify the folding nuclei in proteins: to find residues whose mutations affect the folding rate by changing the transition state stability as strongly as that of the native protein [44,48] (Fig. **5**).

The participation of a residue in the folding nucleus is expressed by the residue's $\phi$ value. $\phi$ is defined as $\Delta\ln k_f/\Delta\ln K$, where $k_f$ is the folding rate constant, $K = k_f/k_u$ is the folding-unfolding equilibrium constant, and $\Delta$ means the mutation-induced shift of the corresponding value. According to the model of a native-like folding nucleus [44,48], $\phi=1$ means that the residue has its native conformation and environment already in the transition state (i.e., that this residue is in the folding nucleus), while $\phi=0$ means that the residue remains unfolded in the transition state. The values $\phi\approx0.5$ (which are observed quite often) are ambiguous: either the residue is at the surface of the nucleus, or it is in one of the alternative nuclei, belonging to a different folding pathway. It is noteworthy that the values $\phi<0$ and $\phi>1$ (which would be inconsistent with the model of a native-like folding nucleus) are extremely rare and never concern a residue with a reliably measured $\Delta\ln K$.

The major assumptions underlying the $\phi$-analysis of the folding nucleus by point mutations [48] are that neither the folding pathway, nor the nucleus, nor the structure of the folded state, nor the unfolded state ensemble is changed as a result of the mutations. Experimentally, this is proved to be usually (there are exceptions [56]) correct when the mutated

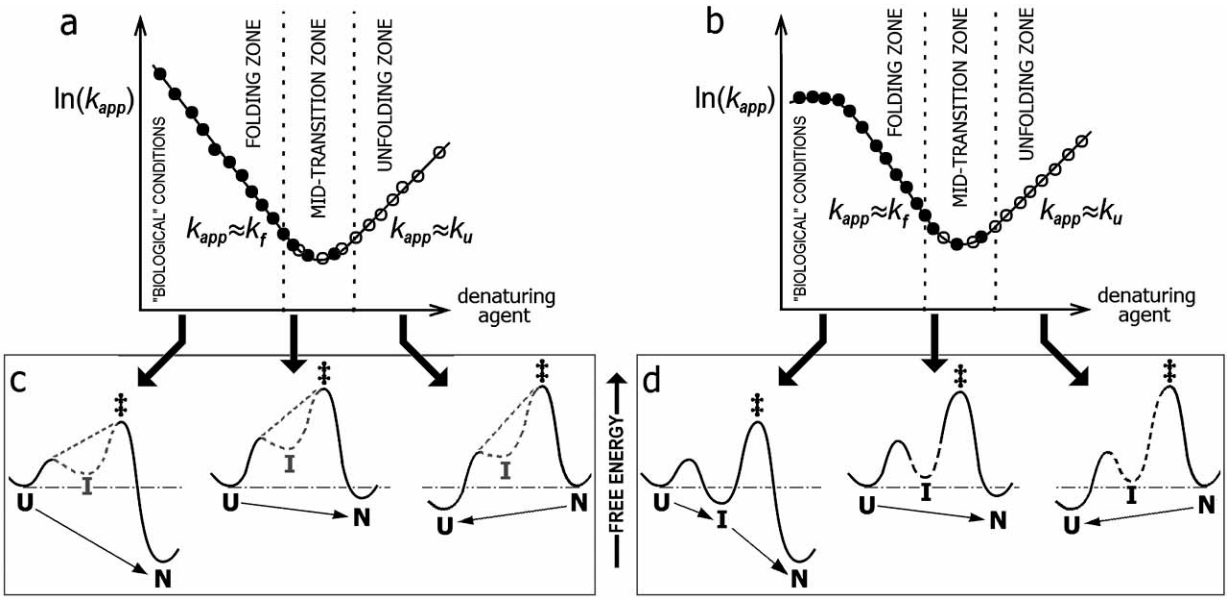


**Fig. (4).** (a, b) Typical appearance of a "chevron plot" presenting an apparent rate of the folding/unfolding process ($k_{app}$) vs. the denaturant concentration (or the temperature). The closed circles correspond to folding which occurs when the protein is transferred from the unfolding to renaturing conditions. The open circles correspond to unfolding that occurs when it is transferred from the native to denaturing conditions. Note that circles of both kinds overlap at mid-transition. (a) Typical plot for a protein having the *two-state* folding throughout the whole range of experimental conditions. For the two-state folding, different characteristics of the protein change with equal rate $k_{app}$, and $k_{app} = k_f + k_u$, where $k_f$ is the folding rate and $k_u$ is the unfolding rate: thus, $k_{app} \approx k_f$ in the folding zone (where $k_f \gg k_u$), $k_{app} \approx k_u$ in the unfolding zone (where $k_f \ll k_u$) and $k_f \approx k_u \approx k_{app}/2$ at the mid-transition [44]. (b) Typical plot for the rate-limiting step of folding and unfolding of a "*multi-state*" folding protein: such a protein has an apparent two-state folding only close to the mid-transition (i.e., the point of thermodynamic equilibrium between the native and denatured states), but not at the "biological" conditions where a multi-state folding occurs (and some state(s) arise(s) and/or decay(s) much faster than the complete transition occurs). The schemes at the bottom of the Figure show the free energy changes along the folding pathways of the two-state (c) and three-state (d) proteins. **N** is the native state, **U** the unfolded, **I** the metastable (molten globule-like) folding intermediate, and ‡ the main, most unstable transition state (as a rule, ‡ is situated directly before the **N** state) [41]. The thin arrows show the directions of the processes, and the dash-dot lines show the free energy levels corresponding to the beginnings of the processes. The dotted lines in (c, d) correspond to those intermediates, which are either invisible in kinetic experiments (this happens when **I** is either unstable relative to the initial state of the process, or is situated after ‡ along the process pathway [11,18]), or absent entirely (the latter possibility is shown by the alternative dotted line).

residue is not larger than the initial one, and when the mutation is not connected with introduction of charges inside the globule; the proof is done by double mutations [57].

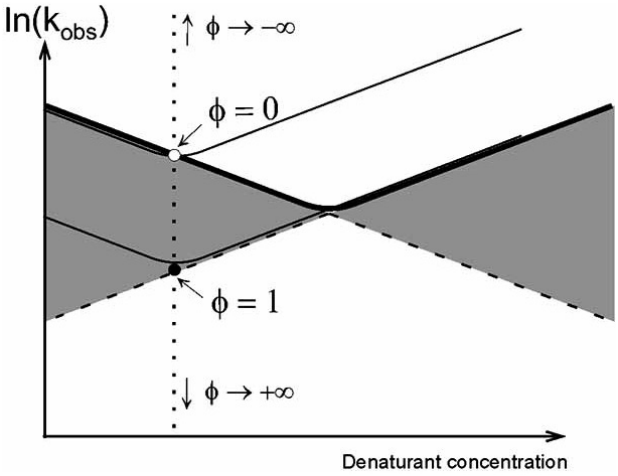


**Fig. (5).** Identification of involvement of a residue into the folding nucleus using site-directed mutations (a scheme). The wild type chevron plot is drawn in bold. The dashed line denotes processes non-observable in experiment (folding at high denaturant concentration and unfolding at low denaturant concentration). (Closed circle) The mid-transition point for chevron plot of a mutant protein if the mutated residue has its native conformation and environment (i.e., its native interactions) already in the folding nucleus; in this case $\phi$=1. (Open circle) The mid-transition point for chevron plot of a mutant protein if the mutated residue remains denatured in the transition state; in this case $\phi$=0. If the mid-transition point moves from open circle to the closed circle, the corresponding $\phi$-value changes from 0 to 1. If the mid-transition point moves up from open circle, then $\phi \to -\infty$; if the mid-transition point moves down from closed circle, then $\phi \to +\infty$. The grey region corresponds to those positions of mid-transition points when $0 \le \phi \le 1$.

The data obtained on the structure of nuclei can be summarized as follows.

(1) In some proteins, the nucleus is situated in the center, in the hydrophobic core [49,58-60]; in some, it is on the boundary of the globule [61,62,63].

(2) In some proteins the nucleus is stabilized by hydrophobic interactions [49,58,62]; in some it includes hydrogen bonds and salt bridges [61,63].

(3) The measurements of the nuclei's solvent-accessible surface area are also rather informative [37]. The ratio of accessible surfaces of the protein's transition and native states can be estimated from the dependence of $k_f$ and $K = k_f/k_u$ (reflecting the transition state and the native protein free energies) on the denaturant concentration $C$: $\beta_T = (\delta \ln k_f/\delta C)/(\delta \ln K/\delta C)$ [64]. When $\beta_T$ is close to 1, the solvent-accessible area of the nucleus is close to that of the native protein; when $\beta_T$ is close to 0, the nucleus is close to that of denatured protein. As a rule, the observed values of $\beta_T$ are close to 0.6 − 0.8 for small proteins. At the same time, the average value of $\phi$ (which shows what fraction of the residue's environment already exists in the transition state) is usually about 0.3 − 0.4. The main difference between $\phi$-values and $\beta_T$ values is that $\phi$-values report on the side chain only, whereas the $\beta_T$ values report on both side chains and main chain. However, specific non-native interactions in the transition state may give rise to $\phi$-values that are negative or larger than unity (cf. [65]).

(4) Proteins with different sequences but similar 3D structures often have similar folding nuclei [61,66-68]. However, there are some exceptions [69] (Fig. **6**). It also has been shown that a circular permutation, which changes the protein topology, sometimes changes [70-72], and sometimes does not change [73], the position of the nucleus within the protein. The observed abundance of $\phi$ values of about 0.5 and the observed sensitivity of nuclei to mutations, together with the results of computer simu-

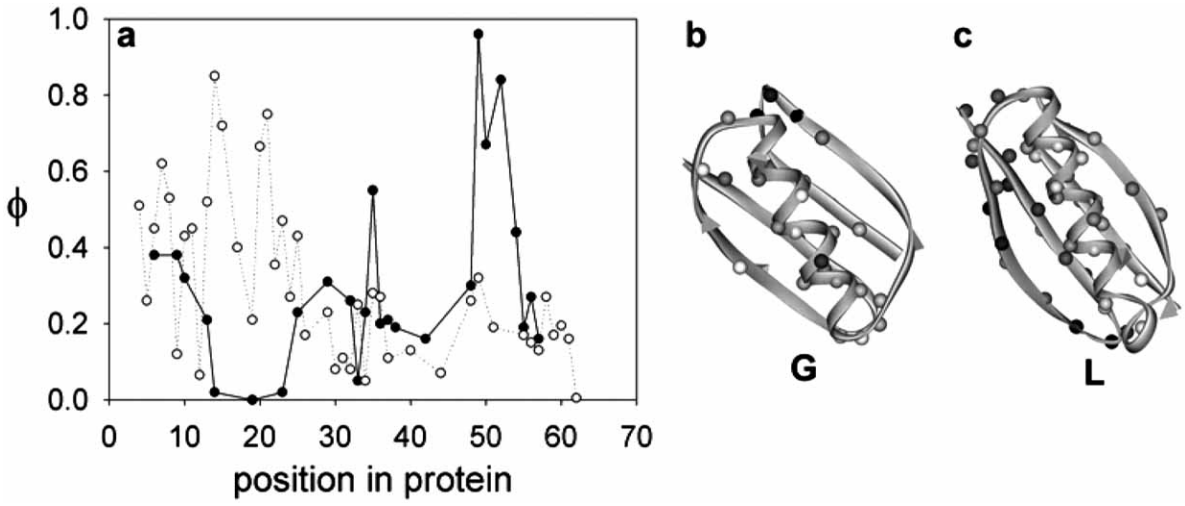


**Fig. (6).** (a) Profiles of experimental $\phi$-values obtained for B1 domains of proteins G (filled circles) and L (open circles). (b,c) Schemes of three-dimensional structures of these proteins [51,52], colored according to the $\phi$ values of the amino-acid residues, from white ($\phi$=0) to black ($\phi$=1). The experimentally studied residues are shown as beads against the background of the native chain fold. Adapted from [53,54]. Sequence identity of G and L proteins is 15% [55].

lations [50,74], led to a conclusion that a "nucleus" is an ensemble of structures rather that a single structure.

Summing up the experimental data, Grantcharova *et al.* conclude that mutations, both artificial and natural, can radically change folding pathways (create and destroy folding intermediates, transforming two- into multi-state folding proteins and *vice versa*, shift the folding nuclei at the opposite side of the molecule, etc.) — without any considerable variation of three-dimensional structures of native proteins [75]. This means that the native structure is a subject of much more severe natural selection than that of the folding nucleus and than folding pathways, – at least when we speak about relatively small proteins, which usually fold much faster than they are synthesized by a ribosome.

## 2. THEORY OF PROTEIN FOLDING

All of the experimental data that we have discussed, although exciting, cannot answer the main question about how a protein manages to find its native, apparently most stable, structure among zillions of others within those minutes or seconds that are assigned for its folding. The explanation for this comes from physical theory.

### 2.1. Solution of the Levinthal Paradox

The difficulty of this "Levinthal problem" is that it cannot be solved by a direct experiment. Indeed, suppose that the protein has some structure that is more stable than the native one but folds very slowly. How can we find it if the protein does not do so itself? Shall we wait for $\sim 10^{10}$ years?

However, is there a real contradiction between "the most stable" and the "rapidly folding" structure? Maybe, the stable structure *automatically* forms a focus for the "rapid" folding pathways, and therefore it is *automatically* capable of fast folding?

Before considering *kinetic* aspects of protein folding, let us recall some basic facts concerning protein *thermodynamics* (as above, we will consider single-domain globular proteins only). This will help us to understand what chains and what folding conditions we have to consider. The facts are as follows:

- Protein folding and unfolding are usually reversible "all-or-none" transitions: only the native and denatured states of the chain are present (close to the denaturation point) in a visible quantity, while the other states are unstable and therefore virtually absent. An "all-or-none" folding phase transition requires the amino acid sequence that provides a large energy gap between the native and the other folds.
- The denatured state, at least that of small proteins unfolded by a strong denaturant, is often the random coil.
- Even under physiological conditions the native state of a protein is only by a few kcal/mol more stable than its unfolded state (and these states have equal stability at mid-transition, naturally).

Thus, to solve the "Levinthal paradox" and to show that the most stable chain fold can be found within a reasonable time, we could, as a first approximation, consider only the rate of the "all-or-none" transition between the coil and the most stable structure. For this, we can consider the case when only one, *the most stable*, chain fold is in (or close to) thermodynamic equilibrium with the coil state, and all other folds of the chain are unstable. Here the analysis is the simplest: it must not consider accumulating intermediates. True, the maximal folding rate is achieved when the native fold is much more stable than the coil (Fig. **4**), and then the observable intermediates often arise. But let us consider the situation when the folding is not the fastest but the simplest…

Since the "all-or-none" transition requires a large energy gap between the most stable structure and the other ones, *we will assume that the considered amino acid sequence provides such a gap.*

We are going to show that the "gap condition" provides a rapid folding pathway to the global energy minimum, to estimate the rate of folding, and to prove that the most stable structure of a normal size domain can fold within seconds or minutes [2,76].

To prove that the most stable chain structure is capable of rapid folding, it is sufficient to prove that at least one rapid folding pathway leads to this structure. Additional pathways can only accelerate the folding since the rates of parallel reactions are additive. [One can imagine water leaking from a full to an empty pool through cracks in the wall between them: when the cracks cannot absorb all the water, each additional crack accelerates filling of the empty pool. And, by definition of the "all-or-none" transition, all semi- and misfolded forms together are too unstable to absorb a significant fraction of the folding chains and trap them.]

A rapid pathway must include not too many steps, but, first of all, it must not require overcoming of a free energy barrier that is too high.

An $L$-residue chain can attain its lowest-energy fold in $L$ steps, each adding one fixed residue to the growing structure (Fig. **7**). *If* the free energy went downhill along the entire pathway, a 100-residue chain would fold in ~ 100 – 1000 ns, since the growth of a structure (e.g., an $\alpha$-helix) by one residue ($\tau$) is known to take a few nanoseconds [77]. Protein folding takes much more than 1 μs only because of the free energy barrier, since most of the folding time is spent on climbing up this barrier and falling back, rather than on moving along the folding pathway.

According to the conventional transition state or Kramer's theories [78,79], characteristic time ($\equiv 1/k$) of the process is estimated as

$$TIME \sim \tau \times \exp(+\Delta F^{\ddagger}/RT) \, . \qquad (3.1)$$

Here $T$ is the temperature, $R$ the gas constant, $\tau$ the time of one step (~ns), and $\Delta F^{\ddagger}$ the free energy of transition state relative to the initial one, i.e., the barrier height.

Our main question is: how high is the free energy barrier $\Delta F^{\ddagger}$ for the pathway leading from the unfolded to the lowest-energy structure?

*If* the fold-stabilizing contacts start to arise only when the chain comes very close to its final structure (that is, if the chain has to lose almost all its entropy *before* the energy starts to decrease), the initial free energy increase would

form a very high free energy barrier (proportional to the *total* chain entropy lost). The Levinthal paradox claiming that the lowest-energy fold cannot be found within any reasonable time, since this involves exhaustive sampling of all chain conformations, originates exactly from this "golf course" picture of the energy landscape (loss of the entire entropy *before* the energy gain).

However, this paradox can be avoided if there is a folding pathway where the entropy decrease is almost immediately covered by the energy decrease (as in the usual first order phase transitions).

Let us consider a *sequential* (Fig. **7**) folding pathway.

At each step of this process, one residue leaves the coil and takes its final position in the lowest-energy 3D structure. This pathway looks a bit artificial, but it is exactly the pathway of unfolding of the lowest-energy structure going in the opposite direction. The detailed balance law [80] reads that direct and reverse reactions must follow the same pathway under the same conditions (and we already agreed to consider the mid-point of the folding-unfolding equilibrium). The advantage of considering this folding scenario is that it obviously exists (though the others are also not excluded); second, it allows us to consider only those residue-residue contacts which exist in the native protein [81]. Thus, we can replace a difficult analysis of folding by a simpler analysis of unfolding, and consider the folding nucleus instead of the nucleus of unfolding: these two nuclei coincide at the thermodynamic mid-transition conditions!

Since, at the equilibrium point, the free energies of the native and unfolded phases are equal, the additional free energy $\Delta F^{\ddagger}$ of the nucleus is due only to the boundary between the native and the unfolded phases. The largest boundary which will be met at the optimal phase transition pathway, i.e., when the boundary between phases moves along the longest axes of the globule (Fig. **7**), includes not more than $\approx L^{2/3}$ out of $L$ residues of the chain. This corresponds to a folding nucleus embracing about half of the globule. The energy of this boundary can be estimated [76] as $\approx \frac{1}{3}\varepsilon L^{2/3}$, where $\varepsilon$ is the protein denaturation energy per residue and $\frac{1}{3}$ is a fraction of the residue's contacts that are lost at a 2-dimensional surface in the 3-dimensional space. The $\varepsilon$ value has been experimentally estimated as $\approx$1 kcal/mol, or $\approx 1.5RT$ at room temperature [12]. Besides, depending on the protein topology and the boundary position, the surface of the nucleus may or may not be covered by unfolded closed loops, whose Flory entropy creates an additional surface tension that adds to the conventional surface energy of the boundary. At the very maximum (when all loops have equal length), this entropic term is

$$T\Delta S^{\ddagger,\mathrm{Flory}} = - L^{2/3} \cdot {}^{1}/_{6} \cdot {}^{5}/_{2} \cdot \ln(3L^{1/3}). \qquad (3.2)$$

Here $L^{2/3}$ is the number of surface amino acid residues, the multiplier $^{1}/_{6}$ reflects that only 1 out of 6 possible directions of the surface residue is consistent with beginning of a loop, the multiplier $^{5}/_{2}$ is used instead of the Flory multiplier $^{3}/_{2}$ because each loop avoids the space occupied by the globule, and $3L^{1/3}$ is the average loop length, when *L/2* non-globular residues are divided into $^{1}/_{6}L^{2/3}$ equal-length loops. However, in a more typical case of random division of *L/2* residues into $^{1}/_{6}L^{2/3}$ loops, the term $-T\Delta S^{\ddagger,\mathrm{Flory}}$ does not exceed [76] $RT \cdot L^{2/3}$ (which is, however, numerically very close to the estimate (3.2) when $L<10^{3}$).

Thus, the free energy barrier for folding (and unfolding) in the mid-transition can be estimated as

$$\Delta F^{\ddagger} \approx (1 \pm 0.5)\, RT \cdot L^{2/3}. \qquad (3.3)$$

$\Delta F^{\ddagger}$ is $\approx 1.5L^{2/3}RT$, when the boundary is densely covered by loops, and $\approx 0.5L^{2/3}RT$, when the boundary is free of them [76]. Since a characteristic time of rearrangement of one residue $\tau$ is $\approx$10 ns [77], it takes $\sim 10\text{ns} \times \exp(1.5L^{2/3})$ to overcome the free energy barrier of nucleation in the first case, and only $\sim 10\text{ns} \times \exp(0.5L^{2/3})$ in the second. This range is exactly consistent with the observed times of protein folding near the mid-transition (Fig. **8**). It has also been estimated that knotting of a 100-residue chain can increase the folding time by no more than two times, and of a 400-residue chain by an order of magnitude at most [82].

The reason for the obtained "non-Levinthal" estimate of achievement of the lowest-energy structure,

$$TIME \sim \exp[(1 \pm 0.5)L^{2/3}] \times 10\text{ns}, \qquad (3.4)$$

is that the entropy decrease is almost immediately compensated for by the energy gain along the sequential folding pathway [84], and the free energy barrier occurs due to the surface effects only.

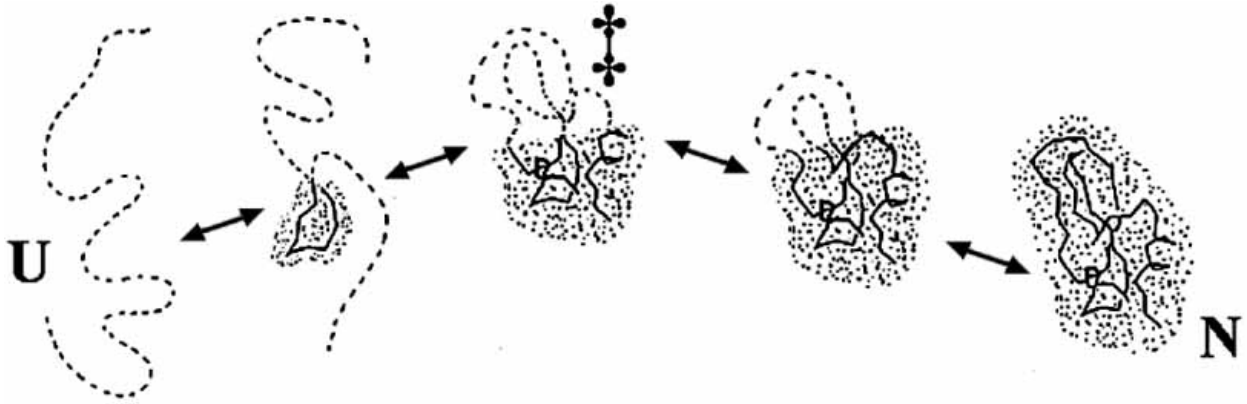


**Fig. (7).** Sequential folding (and unfolding) pathway [76]. U is the unfolded state, N the native state, ‡ the transition state. The folded part (dotted) is native-like. The bold line shows the backbone fixed in this part; the fixed side chains are not shown for the sake of simplicity (the volume that they occupy is dotted). The dashed line shows the unfolded chain.

It is noteworthy that the sequential folding pathway does not require any rearrangement of the dense globular part (which could take a lot of time): all rearrangements occur in the swollen (coil) phase and are therefore rapid.

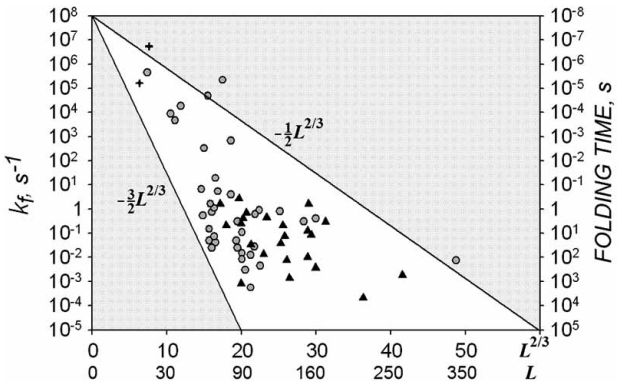


**Fig. (8).** Observed folding rate (and time) at the point of equilibrium between the unfolded and the native states vs. $L^{2/3}$ ($L$ being the number of residues in the chain). The circles and triangles correspond to two-state and multi-state folding proteins, respectively. All these proteins do not contain S-S bonds or large ligands. Symbols + correspond to the α-helical and β-hairpin peptides. The theoretically predicted region of folding rates (white) is between the lines corresponding to **t** = 10ns×exp(0.5$L^{2/3}$) and **t** = 10ns×exp(1.5$L^{2/3}$). The folding time vs. $L^{2/3}$ correlation is 0.65. The folding rate errors do not exceed the size of used symbols. Adapted from [83]. The data ate taken from [108].

Two notes in conclusion of this chapter:

- it is noteworthy that numerous computer simulations of folding never encounter any kinetically inaccessible structures of 3D protein models [74].

- it is remarkable that a recent mathematical estimate [85] of the maximum time necessary to find the most stable structure of a chain molecule in a *d*-dimensional space follows the equation $\ln(TIME) \sim L^{(1-1/d)} \cdot \ln(L)$, in accordance with the above given estimates (3.2) – (3.4), obtained from simple physical considerations.

- computer experiments of Sali and co-workers [86] established that the only requirement for fast folding was a sufficient energy gap between the native state and the next lowest energy state. A successful prediction (Fig. **8**) of the folding rate region by a theory of Finkelstein and Badretdinov [76], that ignores any special construction of the folding nucleus, shows that proteins have only evolved for having sufficient stability of their native folds (and therefore function) and not folding rates (see also a paper by Plaxco and Baker on this [87]).

## 2.2. Theories of Protein Folding Rates

The theory described above not only "demystifies" (as it was written in [74]) "the protein folding problem cast in terms of the Levinthal paradox", but also opens a way to the creation of more detailed theories of protein folding rates.

These theories have to take into account such factors as the non-uniform distribution of strongly and weakly interacting amino acid residues within the globule, the folding pattern ("topology") formed by the native fold of the protein chain and folding under strongly non-equilibrium conditions.

The non-uniform distribution of weak and strong interactions between the nucleus and the remaining part of the protein contributes to $\Delta F^{\ddagger}$ only a term of less than [76] $\sim L^{1/2}RT$. This effect is of secondary importance when the folding takes place close to the mid-transition where $\Delta F^{\ddagger} \sim L^{2/3}RT$. However, the non-uniformity effect seems to be important when the native state is much more stable than the denatured one [74,88-90]. Besides, it is of crucial importance for a theory of those relatively small changes in the folding rates which are caused by mutations of a protein's residues.

When a protein folds under the non-equilibrium conditions, the transition state free energy becomes smaller than in the case of equilibrium conditions (see Fig. **11** below). It was shown that this free energy $\Delta F^{\ddagger}$ scales with the chain length $L$ as $\Delta F^{\ddagger} \sim L^{1/2}RT$ when the native state is moderately more stable than the denatured one [88,90], and as $\Delta F^{\ddagger} \sim (4 \div 6)\ln(L) \cdot RT$ when the native state is much more stable than the denatured one [89,90]. All of these length-based dependencies are consistent with currently available experimental data [91,92].

A recent review [74] of Shakhnovich gives an excellent overview of effects that occur at non-equilibrium folding.

The capillarity model [76] gave rise to the hypothesis that protein folding rates are determined by the average "entropy capacity" (the entropy capacity of an amino acid residue is defined as the number of contacts divided by the number of degrees of freedom; thus, this value is, in a sense, reciprocal to the expected melting temperature) [93]. It has been shown [94] that entropy capacity correlates with folding rates for alpha-helical proteins (correlation coefficient is 0.79) and proteins with mixed (α/β) secondary structure (correlation coefficient is 0.84).

Statistical analysis showed that, among proteins of the same size, α/β proteins have, on the average, a greater number of contacts per residue compared to other protein classes [95]. This difference is due to their more compact (more "spherical") structure of α/β proteins, rather than to tighter packing. The relationship between the average number of contacts per residue and folding rates in globular proteins according to general protein structural class (all-α, all-β, α/β, α+β) has been examined. The analysis demonstrates that α/β proteins have both the greatest number of contacts and the slowest folding rates in comparison to proteins from the other structural classes [95].

The influence of the protein chain topology upon the folding time has been estimated using a "contact order" (CO) parameter [96,97] and other CO-like parameters [98-103]. The CO is equal to the average chain separation of contacting (in the native fold) residues, divided by the chain length. The CO is low for a structure rich in local contact, and high for a structure rich in contacts between remote chain regions. Specifically, CO is low for α-helices and high for β-structure; this may explain the observed 50-fold difference specifically in folding rates of α-helices and β-hairpins [74,104]. However, it has been shown that the fraction of local contacts in protein structure gives a better folding rate prediction [105].

A high CO value reflects the existence of many long closed loops in the fold of a complicated topology; thus, CO is roughly proportional to the "1±0.5" multiplier in Equations (3.3), (3.4), see [106]. When CO is taken into account in addition to the folding rate on the chain length dependence shown in Fig. (**8**), the correlation of theory with experiment rises by an additional 10% and reaches 0.74 [106]. The correlation is nearly the same for two- and multi-state proteins, in contrast to results obtained for "pure" CO in [96], which are good for two-state proteins only. The need for simultaneously taking into account both CO and chain length in protein folding rate estimation also follows from simplified off-lattice folding simulations [107].

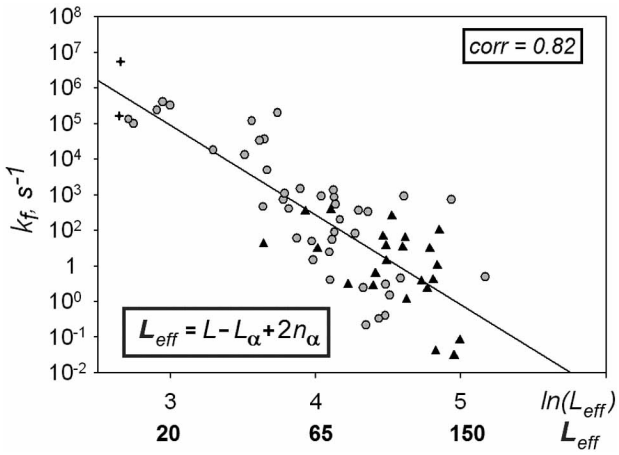


**Fig. (9).** Correlation between observed protein folding rate in denaturant-free water and the number of amino acid residues in the protein chain except for the number of α-helical hydrogen bonds in this chain. The folding rate errors do not exceed the size of used symbols. Adapted from [108].

Interestingly, an even higher correlation (0.82) is observed between the length of an α-helix-free part of a protein chain and the observed protein folding rate in denaturant-free water [108]; in this case, again, the correlation is nearly the same for two- and multi-state proteins (Fig. **9**). It is noteworthy that this highly accurate prediction does not require any knowledge of the protein's tertiary structure: it can be obtained from the amino acid sequence alone, since one can predict [109] α-helices from the sequence.

A growing mass of experimental data has stimulated many studies where experimental folding rates are fitted by a function of secondary structure [110,111] or amino acid sequence, etc. [112-116]. We do not discuss them since they do not follow from any folding theory or folding mechanism, while the interpretation of obtained results suggested by the authors, if any, has obscure physical sense.

A more detailed and physically strict scheme to estimate protein folding rates [117-123] can be obtained from the analysis of the networks of folding pathways, or rather, the networks of pathways of unfolding of native protein structures (Fig. **10**). The free energy of each microstate (a network contains ~$10^6$ microstates) can be estimated [117-121] from the energy of contacts within the folded region and the entropy of the unfolded region (the latter includes the free energy deficit caused by the Flory's entropy of closed unfolded loops as well). In addition to predicting protein folding/unfolding rates at various conditions, a theory based on this scheme is also able to find approximate localization of the protein folding nuclei (see below).

The theoretical results described show, in accordance with experiment, that the most stable protein fold must be found within minutes. They explain also why very large proteins should consist of separately folding domains: otherwise, the chains of more than ~400 residues would fold too

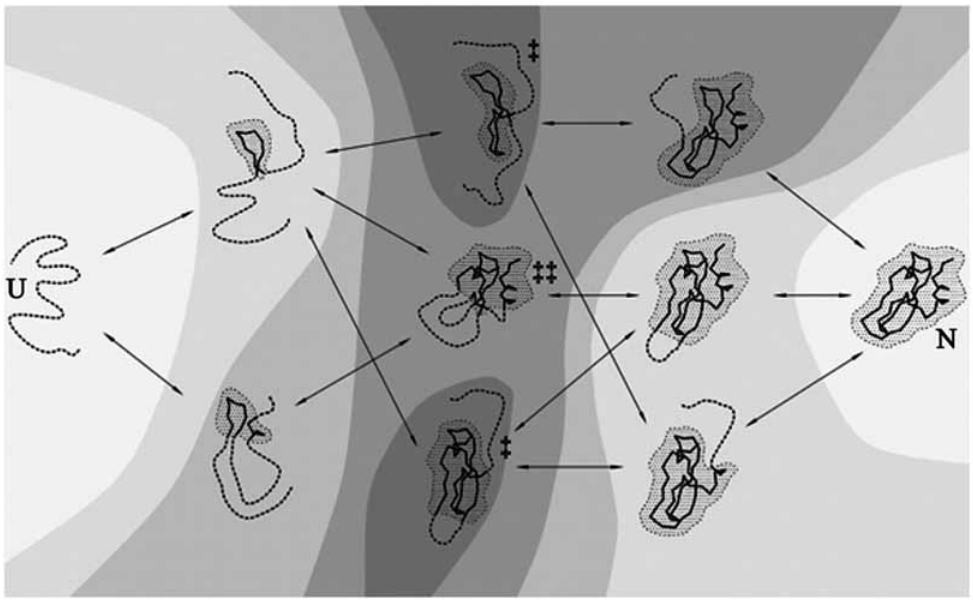


**Fig. (10).** Free energy landscape of a protein chain; the darker the color, the higher the free energy. Arrows show possible transitions between semi-folded microstates (a "transition" is either an addition of one "chain link", i.e., a chain segment including 1 – 5 amino acid residues, to the growing globule or a removal of one chain link from the globule). These transitions are parts of various folding-unfolding pathways, one of which is shown in Fig. (**7**). Transition states are marked with ‡. The main transition state, i.e., the one with the lowest free energy, is marked with ‡‡.

slowly (within days) even when the protein chain topology is simple.

### 2.3. Protein Folding Under Strongly Non-Equilibrium Conditions

So far, we only have considered the folding rate close to mid-transition, where only one (the "native") fold competes with the coil, and all other globular forms, even taken together, are unstable.

If we have an environment that stabilizes the native protein fold, it stabilizes the other globular and semi-folded structures of the chain as well. At first, this stabilization only increases the folding rate (see the rise of the left chevron limb from the mid-transition in Fig. **4**, **5**), since the folding nucleus is also stabilized, and the competing (with the native fold) misfolded structures are still unstable (relative to the initial unfolded state as well) due to the energy gap between them and the native fold (Fig. **11a**).

It has been estimated [76] that each additional 1 kcal/mol in stability of the native (relative to the unfolded) state decreases the barrier height by about ½ kcal/mol (since the critical nucleus includes about half of the protein). This decreased instability of the nucleus accelerates folding at room temperature by about 2.5 times for each additional 1 kcal/mol of native state stability.

However, this acceleration proceeds only up to a certain limit: the maximum folding rate is achieved when the "misfolded" states become as stable as the initial unfolded state, i.e., when the globular folding intermediates become stable (see the plateau at the top of the left chevron limb in Fig. **11b**). The further increase in stability of the folded states leads to a rapid misfolding followed by a relatively slow conversion to the native state (Fig. **11b**). This is a rather complicated process; it includes differences in coming to the native state by the protein main chain and side chains [74], own "folding nucleus" for each step, etc. All of these complications may be considered "artifacts" created by extreme folding conditions (cf. [124]), which are avoided (both in experimental and in theoretical studies) when the protein folding is considered close to mid-transition.

### 2.4. The Search for the Protein Folding Nucleus

The understanding of the nucleation mechanism has a long, contradictory and still unfinished story. The pioneering experimental results [44,48,57] suggested that the folding nucleus is small and specific (i.e., it consists of only a few definite residues). This idea was reinforced by molecular dynamic simulations of protein unfolding (held at a temperature extremely favorable for unfolding) [125] and especially by very accurate and careful analysis of *in silico* folding of simplified lattice protein chains [50], held at a temperature extremely favorable for folding. Later, it was hypothesized that the folding nucleus consists of a small set of highly conserved residues having no obvious functional role [126]. On the contrary, the model of sequential protein folding at the mid-transition temperature [76] suggested to solve the Levinthal paradox led to a conclusion that the nucleus is large (including $\approx$ ⅓ – ½ of the protein molecule) and may be non-specific; for a long time, this was considered a drawback that resulted from oversimplification inherent to this model. However, current estimates of the average $\phi$ values show that $\approx$ ⅓ of the interactions existing in native-state proteins are present in the "folding nucleus" and argue against the existence of a few specific key residues which are *only* important during the kinetics of protein folding [127] (although Fersht and Sato [128] mentioned that the above result was obtained due to undeservedly massive discarding of experimental $\phi$-values with small $\Delta$lnK, i.e., those where experimental errors in $\phi$ values can be more significant [127]). The latest statistical analysis also shows no indication that residues with increased $\phi$-values are significantly more conserved than the others [129]. Ref. [74] argues the opposite, however, and states that the transition state ensemble includes a variety of large nuclei, but that all this ensemble of nuclei includes a small "obligatory" part, common for all nuclei. Thus, theoretical and experimental investigation of the nucleation mechanism is far from being completed.

As regards the theoretical search for folding/unfolding nuclei in proteins, several different approaches have been suggested.

The idea of specific nuclei, reinforced by the lattice simulations of protein folding [50], generated an evolution-

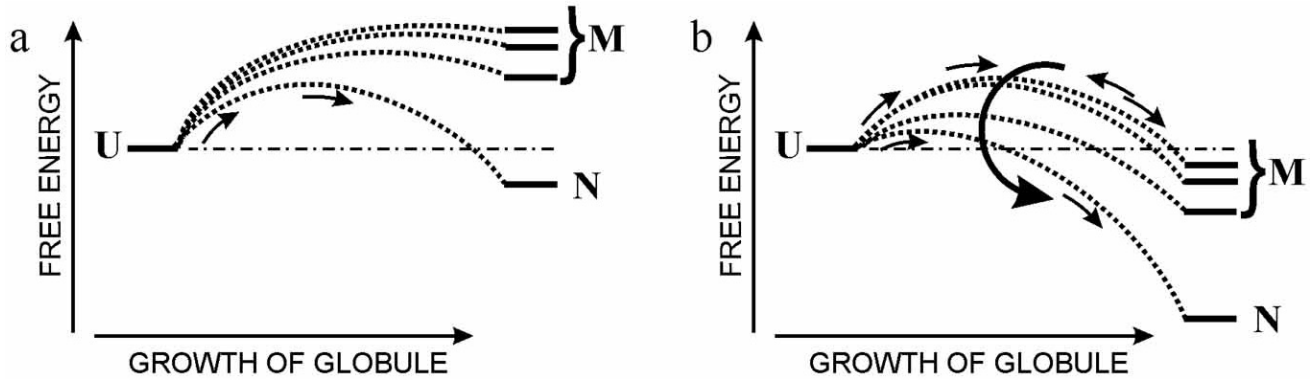


**Fig. (11).** Folding under conditions when (a) the most stable fold **N** is only a little more stable than the unfolded chain **U**, and (b) when **N** is much more stable than **U**, and some mis- or semi-folded structures (shown by numerous free energy levels **M**) are also more stable than **U**. Dotted lines schematically show the free energy changes along the folding pathways leading to different structures; maxima of these lines correspond to the transition states on these pathways. The dash-dot lines show the free energy level of state **U**. Adapted from [76].

ary approach to prediction of the nuclei. It is based on the search for a set of highly conserved residues having no obvious functional role [126,130-132]. It should be mentioned that this approach, at best, can give only the common part of the nuclei existing in homologous proteins. Moreover, some recent observations show that the residue conservation across the homologous proteins correlates with deep immersion into the hydrophobic core of a protein [132] rather than into the folding nucleus [133]. It should be noted that there is some correlation between the nuclei (the regions of high $\phi$ values) and the hydrophobic cores and secondary structures [134-137], but it is rather low on average [133,135].

The most direct approach to the theoretical search for the nucleus generates a plausible transition state using the all-atom molecular dynamic simulations of protein unfolding [125,138-140]. According to these simulations, held for very few small proteins at highly denaturing conditions, the unfolding is hierarchic [141-143] (at least when it occurs far from equilibrium): tertiary interactions break early, whereas secondary structures remain longer. The repeated trajectories show a statistical distribution around the experimentally found transition states and demonstrate a broad ensemble of the TS structures. However, these simulations usually need extremely denaturing conditions (500 K, etc.) to be completed. Therefore, the transition states found for such an extreme *un*folding can be, in principle, rather different from those existing for folding [124]. Recently, however, some molecular dynamic simulations of unfolding of small proteins [144-146] have been performed at more realistic, although also highly denaturating, conditions. They have been performed at temperatures accessible for "wet" experiments (350°K), as well as for simulations at current supercomputers. They gave transition state structures, which are consistent with experiment [146]; however, these simulations take enormous time and can be performed for very small proteins only.

Other approaches are based on a simplified modelling of the protein folding landscapes and trajectories [147-149]. The lattice simulations [150], as well as experiment [56] show that the transition state is an ensemble rather than a single conformation, and that it can be described by an order parameter, such as the fraction of native contacts. The transition state search is based on the projection of the folding trajectories onto a single reaction coordinate (usually, the fraction of the native contacts) and the investigation of barriers at the obtained free-energy profiles [151-153]. Although such a projection is not a rigorous procedure (e.g., the structure which is "nearly native" in terms of contacts can contain an additional knot of the chain and needs complete unfolding before coming to the native state), these studies were able to outline, though crudely, the folding nuclei of some small proteins.

Exciting progress has been made using experimental constraints to obtain the folding nucleus at atomic resolution (or rather, to visualize a possible shape for the folding nucleus that is consistent with the available, but sparse experimental data). Review of these works, excellently done in [74], is outside the scope of our paper.

Important progress is also being made in the prediction of folding nuclei by starting from the known 3D structures of native proteins. This progress is due to the analysis of multi-dimensional networks (see Fig. **10**) of the protein folding-unfolding trajectories done by various algorithms [118,135, 154]. All these approaches (applied also to the studies of the folding rates [76,122], see the end of section 3.2) use different approximations and algorithms, consider only the attractive native interactions (the "Gō model" [155]) to reduce the energy frustrations and heterogeneity of interactions, and model the trade-off between the formation of attractive interactions and the loss of conformational entropy during protein folding. These works also simulate unfolding of known 3D protein structures rather than their folding, but the unfolding is considered close to the mid-transition point, where folding and unfolding pathways coincide according to the detailed balance principle. These works allowed the authors to outline the folding nuclei in more detail. Despite the relative simplicity of these models, they give a promising (~50%) correlation with experiment [156,157,123,119].

It is noteworthy that this correlation is considerably worse than those typical for prediction of protein folding rates. The first obvious reason is that the observed $\phi$ values being predicted are restricted to the narrow region of $0 - 1$ with an experimental error of $\sim\pm0.1$, while the observed folding rates (determined with a relatively small experimental error) belong to the wide range of $10^7$ $s^{-1}$ – $10^{-4}$ $s^{-1}$. A more important reason is that the folding nucleus is not as stable to the action of mutations (and thus to the unavoidable errors in energy estimates used to outline them) as a 3D protein structure (see Fig. **6**), and it would be strange to obtain a perfect prediction of the folding nuclei with the same force fields which are still not able to predict the mutation-stable 3D native structure of a protein [74,158].

## CONCLUDING REMARKS

Actually, protein folding resembles crystallization (where a kind of a Levinthal paradox also exists, since the decrease in the number of configurations during crystallization is huge, but is solved by a mechanism of nucleation) [159]. When crystallization occurs close to the freezing temperature, a perfect, large monocrystal (the lowest-energy structure, an analog of the "native structure") arises, although extremely slowly. As the temperature decreases a little, the monocrystal grows faster; and a further temperature decrease leads to a rapid formation of many chips rather than a perfect large monocrystal [78].

The scheme given above of entropy-by-energy compensation along the folding pathway and the conclusion that it leads to nucleation that can solve the Levinthal paradox are applicable to the formation of the native protein structure not only from the coil but also from the molten globule or from another intermediate. However, for these scenarios, all estimates would be much more cumbersome, while these processes (a variety of which are discussed in Refs. 20, 32, 33, 35, 42, 44, 71, 83, 86–88, 96, 97) do not show (in experiment) drastic advantages in the folding rate. Therefore, here we will not go beyond the simplest case of the coil-to-native globule transition.

Finally, it should be noted that a hierarchic scheme of protein folding [30,31,33,160], as well as many simplified "protein folding funnel" models [161,162] do not solve the

Levinthal paradox since they cannot provide a *simultaneous* [163,164] explanation for all of the three major observations about protein folding: (i) spontaneous, non-assisted folding of a unique native structure within non-astronomical time, (ii) the fact that the same native structure can be achieved at very different conditions (including the thermodynamic mid-transition), from different structural states and with very different folding rates, and (iii) co-existence, to a visible quantity, of only the native and unfolded molecules during folding of single-domain proteins near the thermodynamic mid-transition between the native and denatured states.

On the contrary, a nucleation-condensation mechanism can account for all these major features simultaneously and thus resolves the Levinthal paradox and opens a way to a theory of protein folding nuclei and protein folding/unfolding rates.

## ACKNOWLEDGEMENTS

We are grateful to E. V. Davydova and D. Reifsnyder for assistance in preparation of the paper. We are also grateful to the anonymous referees for careful reading of the manuscript and very helpful comments. This work was supported by the programs "Molecular and cellular biology" and “Fundamental sciences – medicine”, by grants 07-04-01539, 05-04-48750-a of the Russian Foundation for Basic Research, by an INTAS grant (№ 05-1000004-7747), and by an International Research Scholar’s Award to A.V.F. from the Howard Hughes Medical Institute (55005607).

## ABBREVIATIONS

| | | |
|---|---|---|
| GuHCl | = | Guanidine hydrochloride |
| NMR | = | Nuclear magnetic resonance |
| UV CD | = | Ultra violet circular dichroism |
| CO | = | Contact order |
| 3D structure | = | Three dimensional structure |

## REFERENCES

[1] Stryer, L. (**1995**) *Biochemistry*, 4th edn., W.H. Freeman & Co., New York.
[2] Finkelstein, A.V. and Ptitsyn, O.B. (**2002**) *Protein Physics. A Course of Lectures*, Academic Press, An Imprint of Elsevier Science, Amsterdam – Boston – London – N.Y. – Oxford – Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo.
[3] Branden, C. and Tooze, J. (**1999**) *Introduction to Protein Structure*, Garland Publ. Inc., New York.
[4] Uversky, V.N., Gillespie, J.R. and Fink, A.L. (**2000**) *Proteins*, *41*, 415.
[5] Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, H., Wyckoff, H and Phillips, D.C. (**1958**) *Nature*, *181*, 662.
[6] Perutz, M.F., Rossmann, M., Cullis, G.A.F., Muirhead, G., Will, G and North, T. (**1960**) *Nature*, *185*, 416.
[7] Wüthrich, K. (**1986**) *NMR of proteins and nucleic acids*, John Wiley & Sons, New York.
[8] Anfinsen, C.B., Haber, E., Sela, M and White, F.H. (**1961**) *Proc. Natl. Acad. Sci. USA*, *47*, 1309.
[9] Gutte, B. and Merrifield, R.B. (**1969**) *J. Amer. Chem. Soc.*, *91*, 501.
[10] Privalov, P.L. and Khechinashvili, N.N. (**1974**) *J. Mol. Biol.*, *86*, 665.
[11] Creighton, T.E. (**1991**) *Proteins*, 2-nd ed., W. H. Freeman & Co., New York.
[12] Privalov, P.L. (**1979**) *Adv. Protein Chem.*, *33*, 167.
[13] Nojima, H., Ikai, A., Oshima, T. and Noda, H. (**1977**) *J. Mol. Biol.*, *116*, 429.
[14] Privalov, P.L. (**1982**) *Adv. Protein Chem.*, *35*, 1.
[15] Tanford, C. (**1968**) *Adv. Prot. Chem.*, *23*, 121.
[16] Kuwajima, K and Sugai, S. (**1978**) *Biophys. Chem.*, *8*, 247.
[17] Dolgikh, D.A., Abaturov, L.V., Bolotina, I.A., Brazhnikov, E.V., Bychkova, V.E., Bushuev, V.N., Gilmansin, R.I., Lebedev, Yu.O.; Semisotnov, G.V.; Tiktopulo, E.I. and Ptitsyn, O.B. (**1985**) *Eur. Biophys. J.*, *13*, 109.
[18] Ptitsyn, O.B. (**1995**) *Adv. Protein Chem.*, *47*, 83.
[19] Dobson, C.M. (**1994**) *Curr. Biol.*, *4*, 636.
[20] Shakhnovich, E.I. and Finkelstein, A.V. (**1989**) *Biopolymer*, *28*, 1667.
[21] Finkelstein, A.V. and Shakhnovich, E.I. (**1989**) *Biopolymers*, *28*, 1681.
[22] Shakhnovich, E.I. and Gutin, A.M. (**1990**) *Nature*, *346*, 773.
[23] (*a*) Goldstein, R.A., Luthey-Schulten, Z.A and Wolynes, P.G. (**1992**) *Proc. Natl. Acad. Sci. USA*, *89*, 4918; (*b*) Goldstein, R.A.; Luthey-Schulten, Z.A. and Wolynes, P.G. (**1992**) *Proc. Natl. Acad. Sci. USA*, *89*, 9029.
[24] Finkelstein, A.V., Gutin, A.M. and Badretdinov, A.Ya. (**1995**) in *Subcellular Biochemistry* (B. B. Biswas and S. Roy, eds.). vol. 24, pp.1, Plenum Press.
[25] Kolb, V.A., Makeev, E.V. and Spirin, A.S. (**1994**) *EMBO J.*, *13*, 3631.
[26] Ellis, R.J. and Hartl, F.U. (**1999**) *Curr. Opin. Struct. Biol.*, *9*, 102.
[27] Hardesty, B.; Tsalkova, T. and Kramer, G. (**1994**) *Curr. Opin. Struct. Biol.*, *9*, 111.
[28] Fändrich, M., Forge, V., Buder, K., Kittler, M., Dobson, C.M. and Diekmann, S. (**2003**) *Proc. Natl. Acad. Sci. USA*, *100*, 15463.
[29] Lührs, T., Ritter, C., Adrian, M., Riek-Loher, D., Bohrmann, B., Döbeli, H., Schubert, D. and Riek, R. (**2005**) *Proc. Natl. Acad. Sci. USA*, *102*, 17342.
[30] Levinthal, C. (**1968**) *J. Chim. Phys. Chim. Biol.*, *65*, 44.
[31] Phillips, D.C. (**1966**) *Sci. Am.*, *215*, 78.
[32] (*a*) Goldenberg, D.P. and Creighton, T. E. (**1983**) *J. Mol. Biol.*, *165*, 407; (*b*) Goldenberg, D.P. and Creighton, T.E. (**1984**) *J. Mol. Biol.*, *179*, 527.
[33] Ptitsyn, O.B. (**1973**) *Doklady AN SSSR (Moscow)*, *210*, 1213.
[34] Dolgikh, D.A., Gilmanshin, R.I., Brazhnikov, E.V., Bychkova, V.E., Semisotnov, G.V., Venyaminov, S.Yu. and Ptitsyn, O.B. (**1981**) *FEBS Lett.*, *136*, 311.
[35] Dolgikh, D.A., Kolomiets, A.P., Bolotina, I.A. and Ptitsyn, O.B. (**1984**) *FEBS Lett.*, *164*, 88.
[36] Dyson, J.H. and Wright, P.E. (**1996**) *Annu. Rev. Phys. Chem.*, *47*, 369.
[37] Jackson, S.E. (**1998**) *Fold. Des.*, *3*, R81.
[38] Uversky, V.N., Semisotnov, G.V., Pain, R.H. and Ptitsyn, O.B. (**1992**) *FEBS Lett.*, *314*, 89.
[39] Jones, C.M., Henry, E.R., Hu, Y., Chan, C.K., Luck, S.D., Bhuyan, A., Roder, H., Hofrichter, J., Eaton, W.A. (**1993**) *Proc. Natl. Acad. Sci. USA*, *90*, 11860.
[40] Shastry, M.C.R. and Roder, H. (**1998**) *Nature Struct. Biol.*, *5*, 385.
[41] Fersht, A.R. (**1995**) *Curr. Opin. Struct. Biol.*, *5*, 79.
[42] Dobson, C.M. and Karplus, M. (**1999**) *Curr. Opin. Struct. Biol.*, *9*, 92.
[43] Kragelund, B.B., Robinson, C.V., Knudsen, J., Dobson, C.M. and Poulsen, F.M. (**1995**) *Biochemistry*, *34*, 7217.
[44] Matouschek, A., Kellis, J.T.Jr., Serrano, L., Bycroft, M. and Fersht, A.R. (**1990**) *Nature*, *346*, 440.
[45] Maxwell, K.L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M.A., Brown, A.G., Friel, C.T., Hedberg, L., Horng, J.C., Bona, D., Miller, E.J., Vallée-Bélisle, A., Main, E.R., Bemporad, F., Qiu, L., Teilum, K., Vu, N.D., Edwards, A.M., Ruczinski, I., Poulsen, F.M., Kragelund, B.B., Michnick, S.W., Chiti, F., Bai, Y., Hagen, S.J., Serrano, L., Oliveberg, M., Raleigh, D.P., Wittung-Stafshede, P., Radford, S.E., Jackson, S.E., Sosnick, T.R., Marqusee, S., Davidson, A.R. and Plaxco, K.W. (**2005**) *Prot. Sci.*, *13*, 602.
[46] Segawa, S.I. and Sugihara, M. (**1984**) *Biopolymers*, *23*, 2473.
[47] Fersht, A.R. (**1997**) *Curr. Opin. Struct. Biol.*, *7*, 3.
[48] Matouschek, A., Kellis, J.T.Jr., Serrano, L. and Fersht, A.R. (**1989**) *Nature*, *340*, 122.
[49] Itzhaki, L.S., Otzen, D.E. and Fersht, A.R. (**1995**) *J. Mol. Biol.*, *254*, 260.
[50] Abkevich, V.I., Gutin, A.M. and Shakhnovich, E.I. (**1994**) *Biochemistry*, *33*, 10026.

[51] Gallagher, T., Alexander, P., Bryan, P. and Gilliland, G.L. (**1994**) *Biochemistry*, *33*, 4721.
[52] Wikström, M., Drakenberg, T., Forsén, S., Sjöbring, U. and Björck, L. (**1994**) *Biochemistry*, *33*, 14011.
[53] Galzitskaya, O.V. (**2002**) *Mol. Biol. (Moscow)*, *36*, 386.
[54] Finkelstein, A.V. and Galzitskaya, O.V. (**2004**) *Phys. Life Rev.*, *1*, 23.
[55] McCallister, E.L., Alm, E. and Baker, D. (**2000**) *Nat. Struct. Biol.*, *7*, 669.
[56] Burton, R.E., Huang, G.S., Daugherty, M.A., Calderoni, T.L. and Oas, T.G. (**1997**) *Nat. Struct. Biol.*, *4*, 305.
[57] Fersht, A.R., Matouschek, A. and Serrano, L. (**1992**) *J. Mol. Biol.*, *224*, 771.
[58] Fulton, K., Main, E., Daggett, V. and Jackson, S.E. (**1992**) *J. Mol. Biol.*, *291*, 445.
[59] Chiti, F., Taddei, N.,White, P., Bucciantini, M., Magherini, F., Stefani, M. and Dobson, C. (**1999**) *Nat. Struct. Biol.*, *6*, 1005.
[60] Villegas, V., Martinez, J.C., Aviles, F.X. and Serrano, L. (**1998**) *J. Mol. Biol.*, *283*, 1027.
[61] López-Hernándes, E. and Serrano, L. (**1996**) *Fold. Des.*, *1*, 43.
[62] Kragelund, B.B., Osmark, P., Neergaard, T.B., Schiodt, J., Kristiansen, K., Knudsen, J. and Poulsen, F.M. (**1999**) *Nat. Struct. Biol.*, *6*, 594.
[63] Grantcharova, V.P., Riddle, D.S., Santiago, J.V. and Baker, D. (**1998**) *Nat. Struct. Biol.*, *5*, 714.
[64] Matouschek, A., Otzen, D.E., Itzhaki, L.S., Jackson, S.E. and Fersht, A.R. (**1995**) *Biochemistry*, *34*, 13656.
[65] Li, L., Mirny, L.A. and Shakhnovich, E.I. (**2000**) *Nat. Struct. Biol.*, *7*, 336.
[66] Martinez, J.C. and Serrano, L. (**1999**) *Nat. Struct. Biol.*, *6*, 1010.
[67] Riddle, D.S., Grantcharova, V.P., Santiago, J.V., Alm, E., Ruczinski, I. and Baker, D. (**1999**) *Nat. Struct. Biol.*, *6*, 1016.
[68] Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M.A., Jaenicke, R. and Schmid, F.X. (**1998**) *Nat. Struct. Biol.*, *5*, 229.
[69] Steensma, E. and van Mierlo, C.P.M. (**1998**) *J. Mol. Biol.*, *282*, 653.
[70] Viguera, A.R., Serrano, L. and Wilmanns, M. (**1996**) *Nat. Struct. Biol.*, *3*, 874.
[71] Lindberg, M.O., Tangrot, J., Otzen, D.E., Dolgikh, D.A., Finkelstein, A.V. and Oliveberg, M. (**2001**) *J. Mol. Biol.*, *314*, 891.
[72] Lindberg, M.O., Haglund, E., Hubner, I.A., Shakhnovich, E.I. and Oliveberg, M. (**2006**) *Proc. Natl. Acad. Sci. USA*, *103*, 4083.
[73] Otzen, D.E. and Fersht, A. R. (**1998**) *Biochemistry*, *37*, 8139.
[74] Shakhnovich, E. (**2006**) *Chem. Rev.*, *106*, 1559.
[75] Grantcharova, V., Alm, E.J., Baker, D. and Horwich, A.L. (**2001**) *Curr. Opin. Struct. Biol.*, *11*, 70.
[76] Finkelstein, A.V. and Badretdinov, A.Ya. (**1997**) *Fold. Des.*, *2*, 115.
[77] Zana, R. (**1975**) *Biopolymers*, *14*, 2425.
[78] Ubbelohde, A.R. (**1965**) *Melting of Crystal Structure*, Clarendon Press, Oxford.
[79] Moore, J.W. and Pearson, R.G. (**1981**) *Kinetics and Mechanism*, John Wiley & Co., New York.
[80] Lifshiz, E.M. and Pitaevskii, L.P. (**1981**) *Physical Kinetics*, Pergamon, London.
[81] Gō, N. (**1975**) *Int. J. Pept. Prot. Res.*, *7*, 313.
[82] Finkelstein, A.V. and Badretdinov, A.Ya. (**1998**) *Fold. Des.*, *3*, 67.
[83] Finkelstein, A.V., Ivankov, D.N. and Galzitskaya, O.V. (**2005**) *Uspekhi Biol. Khim. (Moscow)*, *45*, 3.
[84] Gō, N. (**1983**) *Ann. Rev. Biophys. Bioeng.*, *12*, 183.
[85] Fu, B. and Wang, W. (**2004**) in *Lecture Notes in Computer Science*, vol. 3142, p. 630.
[86] Sali, A., Shakhnovich, E. and Karplus, M. **(1994)** *J. Mol. Biol.*, *235*, 1614.
[87] Larson, S.M., Ruczinski, I., Davidson, A.R., Baker, D. and Plaxco, K.W. **(2002)** *J. Mol. Biol.*, *316*, 225.
[88] Thirumalai, D. (**1995**) *J. Phys. I* (*Orsay, Fr.*), *5*, 1457.
[89] Gutin, A. M., Abkevich, V.I. and Shakhnovich, E.I. (**1999**) *Phys. Rev. Let.*, *77*, 5433.
[90] Wolynes, P.G. (**1997**) *Proc. Natl. Acad. Sci. USA*, *94*, 6170.
[91] Li, M.S., Klimov, D.K. and Thirumalai, D. (**2004**) *Polymer*, *45*, 573.
[92] Naganathan, A.N. and Munoz, V.M. (**2005**) *J. Am. Chem. Soc.*, *127*, 480.
[93] Galzitskaya, O.V., Surin, A.K. and Nakamura, H. (**2000**) *Prot. Sci.*, *9*, 580.
[94] Galzitskaya, O.V. and Garbuzinskiy, S.O. (**2006**) *Proteins*, *63*, 144.
[95] Galzitskaya, O.V., Reifsnyder, D.C., Bogatyreva, N.S., Ivankov, D.N. and Garbuzynsiy, S.O. **(2007)** *Proteins*, in press.
[96] Plaxco, K.V., Simons, K.T. and Baker, D. (**1998**) *J. Mol. Biol.*, *277*, 985.
[97] Fersht, A.R. (**2000**) *Proc. Natl. Acad. Sci. USA*, *97*, 1525.
[98] Gromiha, M. and Selvaraj, S. (**2001**) *J. Mol. Biol.*, *310*, 27.
[99] Zhou, H. and Zhou, Y. (**2002**) *Biophys. J.*, *82*, 458.
[100] Micheletti, C. (**2003**) *Proteins*, *51*, 74.
[101] Weikl, T.R. and Dill, K.A. (**2003**) *J. Mol. Biol.*, *329*, 585.
[102] Punta, M. and Rost, B. (**2005**) *J. Mol. Biol.*, *348*, 507.
[103] Zhang, L., Li, J., Jiang, Z. and Xia, A. (**2003**) *Polymer*, *44*, 1751.
[104] Eaton, W.A., Muñoz, V., Hagen, S.J., Jas, G.S., Lapidus, L.J., Henry, E.R. and Hofrichter, J. (**2000**) *Annu. Rev. Biophys. Biomol. Struct.*, *29*, 327.
[105] Mirny, L. and Shaknovich, E. (**2001**) *Annu. Rev. Biophys. Biomol. Struct.*, *30*, 361.
[106] Ivankov, D.N., Garbuzynsky, S.O., Alm, E., Plaxco, K.V., Baker. D. and Finkelstein, A.V. (**2003**) *Prot. Sci.*, *12*, 2057.
[107] Koga, N. and Takada, S. (**2001**) *J. Mol. Biol.*, *313*, 171.
[108] Ivankov, D.N. and Finkelstein, A.V. (**2004**) *Proc. Natl. Acad. Sci. USA*, *101*, 8942.
[109] Jones, D.T. (**1999**) *J. Mol. Biol.*, *292*, 195.
[110] Gong, H., Isom, D.G., Srinivasan, R. and Rose, G.D. (**2003**) *J. Mol. BIol.*, *327*, 1149.
[111] Prabhu, N.P. and Bhuyan, A.K. (**2006**) *Biochemistry*, *45*, 3805.
[112] Shao, H., Peng, Y. and Zeng, Z.H. (**2003**) *Prot. Pept. Lett.*, *10*, 277.
[113] Shao, H. and Zeng, Z.-H. (**2003**) *Prot. Pept. Lett.*, *10*, 435.
[114] Kuznetsov, I. and Rackovsky, S. (**2004**) *Proteins*, *54*, 333.
[115] Gromiha, M.M. (**2005**) *J. Chem. Int. Model.*, *45*, 494.
[116] Gromiha, M.M., Thangakani, A.M. and Selvaraj, S. (**2006**) *Nucl. Acids Res.*, *34*, W70.
[117] Ivankov, D.N. and Finkelstein, A.V. (**2001**) *Biochemistry*, *40*, 9957.
[118] Galzitskaya, O.V. and Finkelstein, A.V. (**1999**) *Proc. Natl. Acad. Sci. USA*, *96*, 11299.
[119] Garbuzynskiy, S.O., Finkelstein, A.V. and Galzitskaya, O.V. (**2004**) *J. Mol. Biol.*, *336*, 509.
[120] Garbuzynskiy, S.O.; Finkelstein, A.V. and Galzitskaya, O.V. (**2005**) *Mol. Biol. (Moscow)*, *39*, 1032.
[121] Galzitskaya, O.V., Garbuzynskiy, S.O. and Finkelstein, A.V. (**2005**) *J. Phys. Cond. Matter*, *17*, S1539.
[122] Munoz, V. and Eaton, W. A. (**1999**) *Proc. Natl Acad. Sci. USA*, *96*, 11311.
[123] Alm, E.; Morozov, A.V.; Kortemme, T. and Baker, D. (**2002**) *J. Mol. Biol.*, *322*, 463.
[124] Finkelstein, A.V. (**1997**) *Prot. Eng.*, *10*, 843.
[125] Daggett, V., Li, A., Itzhaki, L.S., Otzen, D.E. and Fersht, A.R. (**1996**) *J. Mol. Biol.*, *257*, 430.
[126] Shakhnovich, E., Abkevich, V. and Ptitsyn, O. (**1996**) *Nature*, *379*, 96.
[127] Sánchez, I.E. and Kiefhaber, T. (**2003**) *J. Mol. Biol.*, *334*, 1077.
[128] Fersht, A.R. and Sato, S. (**2004**) *Proc. Natl Acad. Sci. USA*, *101*, 7976.
[129] Tseng, Y.Y. and Liang, J. (**2004**) *J. Mol. Biol.*, *335*, 869.
[130] Ptitsyn, O.B. (**1998**) *J. Mol. Biol.*, *278*, 655.
[131] Ptitsyn, O.B. and Ting, K.L. (**1999**) *J. Mol. Biol.*, *291*, 671.
[132] Mirny, L.A. and Shakhnovich, E.I. (**1999**) *J. Mol. Biol.*, *291*, 177.
[133] Plaxco, K.W., Larson, S., Ruczinski, I., Riddle, D.S., Thayer, E.C., Buchwitz, B., Davidson, A.R. and Baker, D. (**2000**) *J. Mol. Biol.*, *298*, 303.
[134] Nölting, B. and Andret, K. (**2000**) *Proteins*, *41*, 288.
[135] Galzitskaya, O.V., Skoogarev, A.V., Ivankov, D.N. and Finkelstein, A.V. (**1999**) in *Proceedings of the Pacific Symposium on Biocomputing'2000*, (Altman, R.B.; Dunker, A.K., Hunter, L., Lauderdale, K. and Klein, T.E., Eds.). p. 131, World Scientific, Singapore - New Jersey - London - Hong Kong.
[136] Cota, E., Steward, A., Fowler, S.B. and Clarke, J. (**2001**) *J. Mol. Biol.*, *305*, 1185.
[137] Mirny, L., Shakhnovich, E. (**2001**) *J. Mol. Biol.*, *308*, 123.
[138] Li, A. and Daggett, V. (**1996**) *J. Mol. Biol.*, *257*, 412.
[139] Caflisch, A. and Karplus, M. (**1995**) *J. Mol. Biol.*, *252*, 672.
[140] Brooks III, C.L., Gruebele, M., Onuchic, J.N. and Wolynes, P.G. (**1998**) *Proc. Natl Acad. Sci. USA*, *95*, 11037.
[141] Lazaridis, T. and Karplus, M. (**1997**) *Science*, *278*, 1928.
[142] Isai, J., Levitt, M. and Baker, D. (**1999**) *J. Mol. Biol.*, *291*, 215.

[143] Daggett, V. and Fersht, A.R. (**2003**) *Nat. Rev. Mol. Cell Biol., 4*, 497.
[144] Mayor, U.; Johnson, C.M.; Daggett, V. and Fersht, A.R. (**2000**) *Proc. Natl Acad. Sci. USA, 97*, 13518.
[145] Ferguson, N., Pires, J.R., Toepert, F., Johnson, C. M., Pan, Y.P., Volkmer-Engert, R., Schneider-Mergener, J., Daggett, V., Oschkinat, H. and Fersht, A. (**2001**) *Proc. Natl Acad. Sci. USA, 98*, 13008.
[146] Mayor, U., Guydosh, N.R., Johnson, C.M., Grossman, J.G., Sato, S., Jas, G.S., Freund, S.M.V., Alonso, D.O.V., Daggett, V. and Fersht, A.R. (**2003**) *Nature, 421*, 863.
[147] Onuchic, J.N., Socci, N.D., Luthey-Schulten, Z. and Wolynes, P.G. (**1996**) *Fold. Des., 1*, 441.
[148] Dill, K.A. and Chan, H.S. (**1997**) *Nat. Struct. Biol., 4*, 10.
[149] Shmygelska, A. (**2005**) *Bioinformatics, 21, Suppl. 1*, i394.
[150] Pande, V.S. and Rocksar, D.S. (**1999**) *Proc. Natl Acad. Sci. USA, 96*, 1273.
[151] Shoemaker, B.A., Wang, J. and Wolynes, P.G. (**1999**) *J. Mol. Biol., 287*, 675.
[152] Shoemaker, B.A. and Wolynes, P.G. (**1999**) *J. Mol. Biol., 287*, 657.
[153] Plotkin, S.S. and Onuchic, J.N. (**2000**) *Proc. Natl Acad. Sci. USA, 97*, 6509.
[154] Alm, E. and Baker, D. (**1999**) *Proc. Natl Acad. Sci. USA, 96*, 11305.
[155] Taketomi, H., Ueda, Y. and Gō, N. (**1975**) *Int. J. Pept. Prot. Res., 7*, 445.
[156] Baker, D. (**2000**) *Nature, 405*, 39.
[157] Takada, S. (**1999**) *Proc. Natl Acad. Sci. USA, 96*, 11698.
[158] Krieger, E., Darden, T., Nabuurs, S.B., Finkelstein, A. and Vriend, G. (**2004**) *Proteins, 57*, 678.
[159] Finkelstein, A.V. (**2003**) in *UFJ NATO ASI.* (Barrat, J.-L.; Feigelman, M.; Kurchan, J. and J. Dalibard, Eds.). Session 77, pp. 649. EDP Sciences, Les Ulis – Paris – Cambridge and Springer-Verlag, Berlin – Heidelberg – New York – Hong Kong – London – Milan – Paris – Tokyo.
[160] (*a*) Baldwin, R.L. and Rose, G.D. (**1999**) *Trends Biochem. Sci., 24*, 26; (*b*) Baldwin, R.L. and Rose, G.D. (**1999**) *Trends Biochem. Sci., 24*, 77.
[161] Chan, H.S. and Dill, K.A. (**1998**) *Proteins, 30*, 2.
[162] Bicout, D.J. and Szabo, A. (**2000**) *Prot. Sci., 9*, 452.
[163] Bogatyreva, N.S. and Finkelstein, A.V. (**2001**) *Prot. Eng., 14*, 521.
[164] Finkelstein, A.V. (**2002**) *J. Biomol. Struct. Dyn., 20*, 311.

**Received: November 29, 2006 Revised: July 03, 2007 Accepted: July 03, 2007**