Protein Folding as Flow across a Network of Folding−Unfolding Pathways. 1. The Mid-Transition Case

Dmitry N. Ivankov and Alexei V. Finkelstein*

Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya str., Pushchino, Moscow Region, 142290, Russia

Received: December 26, 2009

Prediction of protein folding rates and folding nuclei is an important problem of protein science. Most of the previously proposed models for protein folding in vitro are based on the nucleation mechanism of this process. Our model considering protein folding as a flow arising in a network of folding−unfolding pathways at a coarse-grained free-energy landscape was described a few years ago, along with an algorithm for calculation of protein folding rates. Here we extend our approach and describe in detail a mathematically strict algorithm for calculating the "folding nuclei", arising as bottlenecks of the flow. Although the proposed physical theory uses no adjustable parameters, its results are in good agreement with experiment. This paper presents (i) the general theory and (ii) the results for the simplest case, i.e., folding/unfolding at the midpoint of thermodynamic equilibrium between the native and unfolded states of a protein; results for "in-water" conditions, i.e., for the case when no denaturant is added and the native state of a protein is much more stable than the unfolded one, will be described in the next paper of the series.

Introduction

Experimental kinetic studies show that in vitro folding of single-domain proteins takes from microseconds[1] to hours:[2] the difference (in orders of magnitude) is like that between the life span of a mouse and the Universe. During the "in-water" (in the absence of denaturant) protein folding one can observe either no visible intermediates[3] or one or even many of them.[2] However, when some admixture of denaturant brings a protein close to the point of thermodynamic equilibrium between the native and denatured (commonly called "unfolded") states, most of single-domain proteins have a simple two-state transition[4,5] and demonstrate no visible intermediates.

Theoretical investigations of protein folding in vitro have shown that models based on a simple nucleation mechanism are capable, in principle, of predicting folding rates[6−9] and outlining residues of key importance for folding (the "folding nucleus"), provided 3D structures of the native proteins are given.[6,9−13] In the past decade, full-atom protein folding simulations were found to be useful for modeling folding and unfolding of small protein domains.[14−18] However, the full-atom simulations still cannot be done for medium-size and large proteins, and the "information about folding pathways, their transition states and intermediates is still coming from simulation of protein unfolding or simulations of folding using simplified model", as noted in ref 19.

This study is the continuation of our previous one, where we introduced a new method, the method of kinetic equations, based on analysis of protein chain "flow" across a network of folding−unfolding pathways. In that paper a strict mathematical algorithm was developed for calculation of protein folding rates. A unique feature of the algorithm among others is that it automatically takes into account all pathways in the network.

The aim of the present study is to describe in detail a strict mathematical algorithm for calculation of not only the rates, but also "ϕ values", which describe participation of various chain links in the folding nucleus of a protein. This algorithm automatically takes into account all possible pathways as well. It uses an analogy between the flow of the protein molecules through the network of protein folding−unfolding pathways and the electric current through the network of electric conductivities. As a consequence, the calculation of ϕ values is analogous, to some extent, to the calculation of heat dissipated on conductivities.

For lucidity, before describing the algorithm for calculation of ϕ values, we first remind the model and the algorithm for calculation of protein folding rates. To explore the effectiveness of both algorithms on the modern set of experimental data, we applied both algorithms to all single-domain proteins, experimentally studied by today. The first paper of the series contains results for the simplest case, i.e., for folding near the point of thermodynamic equilibrium between the native and unfolded states of a protein. The next paper of the series contains results for "in-water" conditions, i.e., for the case when the native state of a protein is much more stable than the unfolded one; the "in-water" conditions are commonly more interesting for experimentalists, since the equilibrium between the native and unfolded states usually requires an admixture of a "nonphysiological" denaturant.

Theory and Methods

Description of the Model. Consider a protein chain having no disulfide bonds. The chain of L amino acid residues is uniformly divided into N links (Figure 1), so that each link consists of $\approx L/N$ residues (for computational reasons, N cannot exceed ≈ 20). Each link has two states: it can be either "folded" or "unfolded"; thus, the N-link chain has 2^N possible microstates. A "microstate" of the chain has a "folded part" (which is assumed to have the same structure as in the native protein) and a fluctuating unfolded part. Contrary to other works,[6,9,11,20] we do not limit the number of unfolded loops protruding from the native-like part of a semifolded protein. The unfolded

* Corresponding author. Tel/Fax: +7(495)514 0218. E-mail: afinkel@vega.protres.ru.

10.1021/jp912186z
Published on Web 05/05/2010

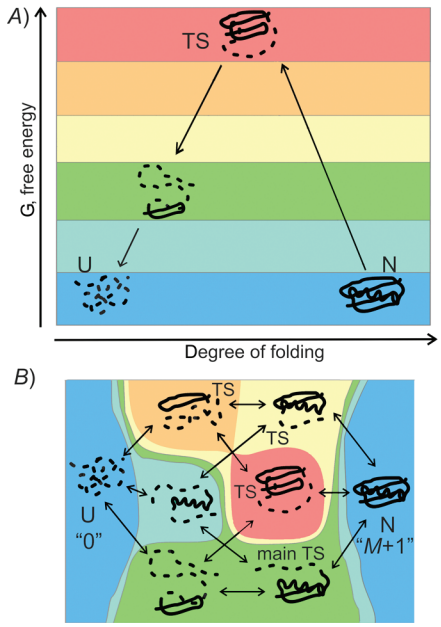

Figure 1. Model of sequential protein unfolding and folding. (A) One possible unfolding pathway of the native state N. Followed in the opposite direction, it is a folding pathway. U, unfolded protein chain. TS, transition state at the given N ↔ U pathway. It corresponds to the top of the free energy barrier at this pathway. The free energy coloring scheme is given. (B) All possible folding−unfolding pathways forming a network. The conformational regions are colored according to their free energies. The state U is numbered here as "0", the state N as "$M + 1$": these indexes, 0 and $M + 1$, are commonly used in the text to denote the most stable unfolded (or nearly unfolded) and folded (or nearly folded) microstates. The "main TS" is TS at the pathway crossing the lowest free-energy barrier. TSs for all other pathways are also shown; they have higher free energies. In the given sketch, the chain consists of only three "links", and thus the whole network of the folding−unfolding pathways consists of $2^3 = 8$ microstates (while in calculations made in the paper we usually have 20 "links" and $2^{20} \approx 1\,000\,000$ microstates). Each arrow corresponds to an elementary transition (see the text).

residues are assumed to have no definite structure and no nonbonded interactions.

We compute free energy G for each microstate, as described in the section "Free Energy of a Microstate" (below). We ascribe the index "0" to an unfolded (or nearly unfolded) microstate of the lowest free energy, and the index "$M + 1$" to a folded (or nearly folded) microstate of the lowest free energy (for details, see "Free Energy of a Microstate"). The other microstates have indices $i = 1, ..., M$. For convenience, we will use dimensionless free energies $\Delta G_i = (G_i - G_0)/RT$ (where R is the gas constant, T is temperature, and $i = 0, 1, ..., M, M + 1$). Thus, $\Delta G_0 \equiv 0$.

Microstates i and j ($i, j = 0, 1, ..., M, M + 1$) are connected by an "allowed elementary transition", if one of them can be obtained from the other by unfolding or folding of one chain link (Figure 1 *B*).

Let k_{ij} be the rate of direct transition from a microstate i to a microstate j (and k_{ji} from j to i transition). If these microstates are connected by an "allowed elementary transition", $k_{ij} \neq 0$; if they are not connected, $k_{ij} = 0$. In particular, $k_{0,M+1} = k_{M+1,0} = 0$ (since protein unfolding requires unfolding of many links), and $k_{ii} \equiv 0$ for all i. For details of computation of k_{ij} from ΔG_i, ΔG_j values, see the section "Rate of Elementary Transition".

The model described above gives a possibility to compute, with certain simplifications, a flow of molecules across a network of folding−unfolding pathways. In this model, the major simplification is replacement of a detailed simulation of infinitely diverse chain movements by solution of a finite set of kinetic equations. Another important simplification consists in ignoring the non-native interactions, which, being weak but numerous, can create metastable or even stable intermediates, like a molten globule.

Algorithm for Calculation of the Folding and Unfolding Rates. Population n_i of a microstate i ($i = 0, ..., M + 1$), i.e., the number of molecules having this microstate, changes with time according to the common kinetic equation:

$$\frac{dn_i}{dt} = -\sum_{\substack{j=0,\\ j\neq i}}^{M+1} k_{ij} n_i + \sum_{\substack{j=0,\\ j\neq i}}^{M+1} k_{ji} n_j \qquad (i = 0, ..., M + 1) \tag{1}$$

The first sum describes a flow from i, the second to i. Note that $\Sigma_{i=0}^{M+1}(dn_i/dt) = 0$, and therefore, $\Sigma_{i=0}^{M+1} n_i$ is constant. Since two microstates, 0 and $M + 1$, are assumed to be special (namely, these are two stable states of the system) and $k_{0,M+1} = k_{M+1,0} = 0$, the set of eqs 1 can be presented as

$$\begin{cases} \dfrac{dn_0}{dt} = -(\sum_{j=1}^{M} k_{0j}) n_0 + \sum_{j=1}^{M} k_{j0} n_j \\ \dfrac{dn_i}{dt} = -(\sum_{\substack{j=0,\\ j\neq i}}^{M+1} k_{ij}) n_i + \sum_{\substack{j=1,\\ j\neq i}}^{M} k_{ji} n_j + \\ \qquad k_{0i} n_0 + k_{M+1,i} n_{M+1} \qquad (i = 1,...,M) \\ \dfrac{dn_{M+1}}{dt} = -(\sum_{j=1}^{M} k_{M+1,j}) n_{M+1} + \sum_{j=1}^{M} k_{j,M+1} n_j \end{cases} \tag{2}$$

Introducing matrix **A** with elements

$$\mathbf{A}_{ij} = \begin{cases} \sum_{\substack{p=0,\\ p\neq i}}^{M+1} k_{ip}, i = j \\ -k_{ji}, i \neq j \end{cases}$$

($i, j = 1, ..., M$), vectors **n**, **B**, **B***, **C**, **C*** (with elements n_i, $B_i = k_{0i}$, $B_i^* = k_{i0}$, $C_i = k_{M+1,i}$ and $C_i^* = k_{i,M+1}$, respectively; $i = 1, ..., M$) and constants $b = \Sigma_{i=1}^{M} k_{0i} \equiv \Sigma_{i=1}^{M} B_i$, $c = \Sigma_{i=1}^{M} k_{M+1,i} \equiv \Sigma_{i=1}^{M} C_i$, we obtain the set of eqs 2 in a more compact matrix form:

$$\begin{cases} \dfrac{dn_0}{dt} = -bn_0 + \mathbf{B}^*\mathbf{n} \\ \dfrac{d\mathbf{n}}{dt} = -\mathbf{A}\mathbf{n} + \mathbf{B}n_0 + \mathbf{C}n_{M+1} \\ \dfrac{dn_{M+1}}{dt} = -cn_{M+1} + \mathbf{C}^*\mathbf{n} \end{cases} \tag{3}$$

Using a quasi-stationary approximation[21] $dn_i/dt = 0$ for each unstable, i.e., poorly populated and thus slowly changing its population microstate $i = 1, ..., M$, we obtain a vector equation $d\mathbf{n}/dt = \mathbf{0}$, i.e., an equation set

$$-\mathbf{A}\mathbf{n} + \mathbf{B}n_0 + \mathbf{C}n_{M+1} = \mathbf{0} \tag{4}$$

Its solution,

$$\mathbf{n} = (\mathbf{A}^{-1}\mathbf{B})n_0 + (\mathbf{A}^{-1}\mathbf{C})n_{M+1} \quad (5)$$

is expressed via a set of rate constants k_{ij} ($i, j = 0, 1, ..., M, M + 1$) and populations of only two microstates, 0 and $M + 1$. Then,

$$\frac{dn_0}{dt} = -(b - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B})n_0 + (\mathbf{B}^*\mathbf{A}^{-1}\mathbf{C})n_{M+1} \quad (6)$$

$$\frac{dn_{M+1}}{dt} = (\mathbf{C}^*\mathbf{A}^{-1}\mathbf{B})n_0 - (c - \mathbf{C}^*\mathbf{A}^{-1}\mathbf{C})n_{M+1} \quad (7)$$

Since $\sum_{i=0}^{M+1}(dn_i/dt) = 0$ and $dn_i/dt = 0$ for $i = 1, ..., M$, we have $dn_0/dt = -dn_{M+1}/dt$.

Thus, in a quasi-stationary approximation we have an "all-or-none" transition between the states 0 and $M + 1$,

$$\frac{dn_0}{dt} = -k_f n_0 + k_u n_{M+1} = -\frac{dn_{M+1}}{dt} \quad (8)$$

where the resulting folding rate constant k_f and the resulting unfolding rate constant k_u are, respectively,

$$k_f = \mathbf{C}^*\mathbf{A}^{-1}\mathbf{B} \quad (9)$$

(note that comparing eqs 6 and 7 one obtains $b - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B} = \mathbf{C}^*\mathbf{A}^{-1}\mathbf{B}$) and

$$k_u = \mathbf{B}^*\mathbf{A}^{-1}\mathbf{C} \quad (10)$$

(note that comparison of eqs 6 and 7 gives $c - \mathbf{C}^*\mathbf{A}^{-1}\mathbf{C} = \mathbf{B}^*\mathbf{A}^{-1}\mathbf{C}$).

The overall flow consists of the "folding flow" with the rate constant k_f and the "unfolding flow" (headed in the opposite direction) with the rate constant k_u.

The rate constant k_f is a dot product of vectors $\mathbf{C}^*$ and

$$\mathbf{n}^{(f)} = \mathbf{A}^{-1}\mathbf{B} \quad (11)$$

we find the vector $\mathbf{n}^{(f)}$ from the matrix equation $\mathbf{A}\mathbf{n}^{(f)} = \mathbf{B}$ numerically by the Zeidel's iteration method.[22] The matrix eq 11 is a special case of eq 5 with $n_0 = 1$ and $n_{M+1} = 0$ (a case of "pure folding").

Similarly, the rate constant k_u is a dot product of vectors $\mathbf{B}^*$ and

$$\mathbf{n}^{(u)} = \mathbf{A}^{-1}\mathbf{C} \quad (12)$$

where $\mathbf{n}^{(u)}$ is found, also by the Zeidel's iteration method,[22] from the matrix equation $\mathbf{A}\mathbf{n}^{(u)} = \mathbf{C}$, which corresponds to eq 5 with $n_{M+1} = 1$ and $n_0 = 0$ (a case of "pure unfolding").

It should be noted that our method, unlike, for example, a method of dynamic programming,[11,20] takes into consideration the entire variety of folding−unfolding pathways. This means that it considers not only the folding pathways where the number of folded links grows permanently but also those with the number of folded links alternatively growing and decreasing.

Algorithm for Calculation of the Folding Nuclei. The folding nucleus is conventionally[23,24] determined as a set of residues forming a folded part of the transition state of the protein's folding. Mutations of these residues have equal or nearly equal influence on stability of the transition state and on that of the folded protein. The ϕ value, reflecting participation of a mutated residue in the folding nucleus, is calculated[23] as

$$\phi = \frac{\delta \ln k_f}{\delta \ln K} \quad (13)$$

where $K = k_f/k_u$ is a constant of equilibrium between the folded and unfolded states of the protein, and δ denotes a mutation-induced change of the value.

In the language of flows, the ϕ value can be computed as follows.

An elementary flow from the microstate i to the microstate j is

$$I_{ij} = k_{ij}n_i - k_{ji}n_j = -I_{ji} \qquad (i,j = 0,..., M + 1, i \neq j) \quad (14)$$

It should be noted that the rate constants of direct and reverse reactions, k_{ij} and k_{ji}, are connected with free energies G_i, G_j of the microstates i, j by well-known[21] relationship

$$k_{ij}\exp(-G_i/RT) = k_{ji}\exp(-G_j/RT) \quad (15)$$

this is the case because equilibrium populations of the states i, j are proportional to Boltzmann factors $\exp(-G_i/RT)$ and $\exp(-G_j/RT)$, respectively, and there is no flow in the equilibrium. Thus, using $\Delta G_i = (G_i - G_0)/RT$ and $\Delta G_j = (G_j - G_0)/RT$, we can rewrite eq 14 as

$$I_{ij} = [k_{ij}\exp(-\Delta G_i)]\cdot\left[\frac{n_i}{\exp(-\Delta G_i)}\right] - [k_{ji}\exp(-\Delta G_j)]\cdot\left[\frac{n_j}{\exp(-\Delta G_j)}\right] = \sigma_{ij}[U_i - U_j] = -I_{ji} \quad (16)$$

The electric current equation has a direct analogy with this equation; we see that the flow I_{ij} plays the role of current from node i to node j, that

$$\sigma_{ij} \equiv k_{ij}\exp(-\Delta G_i) = \sigma_{ji} \equiv k_{ji}\exp(-\Delta G_j) \quad (17)$$

plays the role of conductivity connecting these nodes, and that

$$U_i \equiv \frac{n_i}{\exp(-\Delta G_i)} \quad (18)$$

plays the role of voltage at node i (thus, U_i can be obtained from ΔG_i and n_i, obtained from eq 5) This "flow-current" analogy can be extended by two further steps: (i) n_i plays the role of a charge at a condenser connecting node i with "earth" having zero voltage, and $\exp(-\Delta G_i)$ the role of capacity of this condenser; (ii) the quasi-stationary approximation ($dn_i/dt = 0$

for unstable intermediate microstates 1, ..., M) is an analogue of Kirchhoff's current law ($\Sigma_{j=0}^{M+1} I_{ij} = 0$ for the nodes 1, ..., M).

By analogy to thermal power dissipated at conductivity ij, we can introduce a value,

$$W_{ij} = (U_i - U_j)^2 \sigma_{ij} = (U_i - U_j) I_{ij} \quad (i, j = 0, ..., M+1, i \neq j) \quad (19)$$

then $W_{ij} = W_{ji}$, and

$$W \equiv \sum_{i=0}^{M+1} \sum_{j>i}^{M+1} W_{ij} \quad (20)$$

plays the role of thermal power dissipated by the total system.

On the other hand, from the same "current analogy",

$$W = (U_0 - U_{M+1})\left(-\frac{dn_0}{dt}\right) \quad (21)$$

where $-dn_0/dt = I_{0\to M+1} = dn_{M+1}/dt$ is a total current through the system in a quasi-stationary approximation, and $U_0 = n_0/\exp(-\Delta G_0) = n_0$ (since $\Delta G_0 \equiv 0$). Note that $U_{M+1} = 0$ for the "folding flow", since $n_{M+1} = 0$ in this case. Since the resulting "folding flow" satisfies the equation $-dn_0/dt = k_f n_0$ (cf. eq 8), $W = k_f n_0^2 = k_f U_0^2$ in this case; and, as we see,

$$k_f = \frac{W}{n_0^2} \quad (22)$$

plays the role of the total system's conductivity.

Suppose we mutate some residue and thereby change G_i values by δG_i ($i = 0, ..., M+1$). Then $\Delta G_i \to \Delta G_i + \delta \Delta G_i$ (where $\delta \Delta G_i = \delta G_i - \delta G_0$, and hence $\delta \Delta G_0 \equiv 0$). Since $k_f = W/n_0^2$, $K = k_f/k_u = \exp(-\Delta G_{M+1}) \equiv \exp[-(G_{M+1} - G_0)/RT]$, and n_0 is not changed by mutation, eq 13 obtains the form

$$\phi = -\frac{\delta \ln k_f}{\delta \Delta G_{M+1}} = -\frac{1}{\delta \Delta G_{M+1}} \frac{\delta W}{W} \quad (23)$$

Using eqs 19, 20, and 16, we have

$$\delta W = \sum_{i=0}^{M+1} \sum_{j>i}^{M+1} \delta W_{ij} = \frac{1}{2} \sum_{i=0}^{M+1} \sum_{\substack{j=0, \\ j\neq i}}^{M+1} \delta[(U_i - U_j)^2 \sigma_{ij}] = \sum_{i=0}^{M+1} \sum_{\substack{j=0, \\ j\neq i}}^{M+1} \left[I_{ij}(\delta U_i - \delta U_j) + \frac{1}{2}(U_i - U_j)^2 \delta \sigma_{ij} \right] \quad (24)$$

Since $I_{ij} = -I_{ji}$, the sum of the first terms in eq 24 can be rearranged as

$$\sum_{i=0}^{M+1} \sum_{\substack{j=0, \\ j\neq i}}^{M+1} I_{ij}(\delta U_i - \delta U_j) = 2 \sum_{i=0}^{M+1} \sum_{\substack{j=0, \\ j\neq i}}^{M+1} I_{ij} \delta U_i = 2 \sum_{i=0}^{M+1} \delta U_i \sum_{\substack{j=0, \\ j\neq i}}^{M+1} I_{ij} = 2 \sum_{i=0}^{M+1} \delta U_i \left(-\frac{dn_i}{dt}\right)$$

Since all $dn_i/dt = 0$ at $i \neq 0, M+1$ in the used quasi-stationary approximation, while $\delta U_{M+1} = 0$ (because $U_{M+1} \equiv 0$ for the "folding flow"; see eq 18, where $n_{M+1} \equiv 0$ for this case) and $\delta U_0 = 0$ (see eq 18, where $\delta \Delta G_0 \equiv 0$ and n_0 is also unchanged by mutation), the sum of the first terms in eq 24 turns to zero, and

$$\delta W = \sum_{i=0}^{M+1} \sum_{j>i}^{M+1} (U_i - U_j)^2 \delta \sigma_{ij} = \sum_{i=0}^{M+1} \sum_{j>i}^{M+1} W_{ij} \delta(\ln \sigma_{ij}) \quad (25)$$

(see eq 19). Since $\delta(\ln \sigma_{ij}) = \delta(\ln k_{ij} - \Delta G_i)$ (cf. eq 17), the value of ϕ (cf. eq 23) can be presented as

$$\phi = \sum_{i=0}^{M+1} \sum_{j>i}^{M+1} \frac{W_{ij}}{W} \times \frac{\delta \Delta G_i - \delta \ln k_{ij}}{\delta \Delta G_{M+1}} \quad (26)$$

It remains to connect $\delta \ln k_{ij}$ with $\delta \Delta G$ terms. From Metropolis approximation of the folding rates (see eq 36 below) we can have

$$\delta \ln k_{ij} = \delta \ln[\min(1, \exp(\Delta G_i - \Delta G_j))] = (\delta \Delta G_i - \delta \Delta G_j) \frac{1 - \mathrm{sign}(\Delta G_i - \Delta G_j)}{2} \quad (27)$$

where

$$\mathrm{sign}(\Delta G_i - \Delta G_j) = \begin{cases} 1, & \Delta G_i > \Delta G_j \\ 0, & \Delta G_i = \Delta G_j \\ -1, & \Delta G_i < \Delta G_j \end{cases}$$

Therefore, $\delta(\Delta G_i - \ln k_{ij}) = {}^1/_2(\delta \Delta G_i(1 + \mathrm{sign}(\Delta G_i - \Delta G_j)) + \delta \Delta G_j(1 + \mathrm{sign}(\Delta G_j - \Delta G_i)))$. Then

$$\phi = \sum_{i=0}^{M+1} \sum_{\substack{j=0, \\ j\neq i}}^{M+1} \frac{\delta \Delta G_i}{\delta \Delta G_{M+1}} \frac{W_{ij}}{W} \frac{(1 + \mathrm{sign}(\Delta G_i - \Delta G_j))}{2} \quad (28)$$

Applying this equation to mutation of the residue α (and thus to $\phi = \phi_\alpha$), one can write

$$\phi_\alpha = \sum_{i=0}^{M+1} \left[\frac{\delta \Delta G_i}{\delta \Delta G_{M+1}}\right]_\alpha \times P_i \quad (29)$$

The value $[\delta \Delta G_i/\delta \Delta G_{M+1}]_\alpha$, which depends on the inserted mutation, can be calculated (see next chapter) from the mutation-

caused differences in numbers of contacts (ν) in the microstates i and $M + 1$:

$$\left[\frac{\delta\Delta G_i}{\delta\Delta G_{M+1}}\right]_\alpha = \frac{[\delta\nu_i]_\alpha}{[\delta\nu_{M+1}]_\alpha} \tag{30}$$

The value

$$P_i = \sum_{\substack{j=0,\\ j\neq i}}^{M+1} \frac{W_{ij}}{W} \frac{1 + \mathrm{sign}(\Delta G_i - \Delta G_j)}{2} \tag{31}$$

on the contrary, does not depend on the mutated residue α (it depends only on flows passing the microstate i and free energy differences between this microstate and the others); the values W and W_{ij} ($i, j = 0, ..., M + 1$, $i \neq j$) can be found from their above given definitions (see eqs 19, 20, and 16−18), using $n_0 = 1$ and $\mathbf{n}^{(f)}$ vectors calculated from eq 11.

Continuing the "flow-current" analogy, one can say that the highest ϕ_α value is achieved when we subject to mutation a residue α that gives the maximal change of $\delta\Delta G$ in the microstate i having a higher G as compared with its neighbors j connected with i by the "hottest" (producing the maximal "heat" W_{ij}) conductors.

However, the sum presented in eq 29 shows that the ϕ_α value is not, generally speaking, produced solely by the microstate i, and that a concept of "transition state" itself is a kind of simplification, when a more realistic picture of the pathway network is considered. Actually, the value of ϕ_α depends on the effect of mutated residue α on the ensemble of transition states ("hot connections") in the network of folding−unfolding pathways.

Free Energy of a Microstate. Free energy G_i of a microstate i is calculated from the protein chain structure in this microstate. The surrounding water is taken into account implicitly, via mean-force potentials contributing to the "interaction energy" E_i. Thus,

$$G_i = E_i - TS_i \tag{32}$$

where E_i is the "energy" of protein's nonbonded interactions in the microstate i (which implicitly, due to the mean-force potentials, includes energy and entropy of interactions with solvent as well), S_i is the protein's conformational entropy of this microstate, T is temperature.

Similar to our previous paper,[11] here we use Go-like potentials,[25] i.e., take into account only those inter-residue interactions, which exist in the native structure. Thus, energy E_i is proportional to the number of contacts ν_i between atoms of those amino acid residues that remain in their native positions in the microstate i:

$$E_i = \varepsilon\nu_i \tag{33}$$

(the Go potentials use the same value ε for energy of each native atom−atom contact). Atoms are considered to "be in contact" if they belong to different and non-neighboring chain residues (since the atom−atom contacts within a residue or between chain neighbors remain in the unfolded chain in nearly the same amount as in the folded one), and if the distance between these atoms does not exceed the sum of their contact radii. Following the study,[20] we added hydrogen atoms to all protein structures where such atoms have not been resolved; a program YASARA[26] was used to this purpose. Hydrogen atoms are assigned contact radii of 2 Å, all non-hydrogen atoms of 3 Å.

The entropy S_i is calculated as

$$S_i = s\cdot(L - m_i) + \sum_{\text{loops}} S_{\text{loop}} \tag{34}$$

where m_i is the number of residues fixed in the native conformation in the microstate i, $s = 2.3R$ is the experimentally measured average entropy lost by a residue after protein melting,[27] and S_{loop} is the entropy of a loop, protruding from the globular part of the microstate; it is estimated as in refs 7 and 11:

$$S_{\text{loop}} = -\frac{5}{2}R\ln|\alpha - \beta| - \frac{3}{2}\frac{R(r_{\alpha\beta}{}^2 - r_{12}{}^2)}{2Ar_{12}|\alpha - \beta|} \tag{35}$$

Here $r_{\alpha\beta}$ is the distance between C_α atoms of the residues α and β, which are situated before and after the unfolded loop, $r_{12} = 3.8$ Å is the distance between C_α atoms of consecutive residues, R is the gas constant, and $A = 20$ Å is the persistent length of a polypeptide chain known from experiment.[28] (It should be noted that the value of S_{loop} is, actually, determined by the first, logarithmic term, while the second one is relatively small.) The use of the factor 5/2 as a multiplier of the logarithmic term instead of 3/2, generally typical of 3D coils,[28] reflects the fact that in semimolten protein the loop is situated above the globule's surface.[7]

Rate of Elementary Transition. If an elementary transition from the microstate i to the microstate j is "allowed", i.e., if one of the microstates can be obtained from another by folding or unfolding of one chain link, the rate of the transition is determined according to Metropolis:[29]

$$k_{ij} = k_0 \min(1,\exp(\Delta G_i - \Delta G_j)) \qquad (i, j = 0, ..., M + 1; i \neq j) \tag{36}$$

where k_0 is the rate of downhill transition; for transitions which are not allowed, $k_{ij} = 0$. We take $k_0 = k^*/(L/N)$, where $k^* = 10^8$ s^{-1} is the measured rate of reordering of one residue[30] and L/N is the number of residues in a chain link. The term L/N reflects the influence of the link size on the rate of elementary transition. This estimate follows from the simplest assumption that reordering of L/N residues lasts L/N times longer than reordering of one residue. Of course, this is a rough assumption, and the use of one and the same k_0 for links having different environment, shape and flexibility is even rougher. An excuse is that here we search for the lowest free-energy barriers between the folded and unfolded states of proteins rather than make a detailed simulation of conformational changes.

Chain Division into "Links". As it has been already mentioned, we cannot use chains of more than ≈20 links in our computations.

Let L, "the number of residues in a folding protein chain", be the total number of protein chain residues minus the number of residues in its unresolved N- and C-terminal tails. If unresolved N-terminal tail contains l_0 residues, $l'_1 = l_0 + 1$ is the chain number of the first residue of the first link. To divide a sequence of L residues into $N < L$ links in a uniform way, the chain number l''_n of the last residue of link n ($1 \leq n \leq N$) is

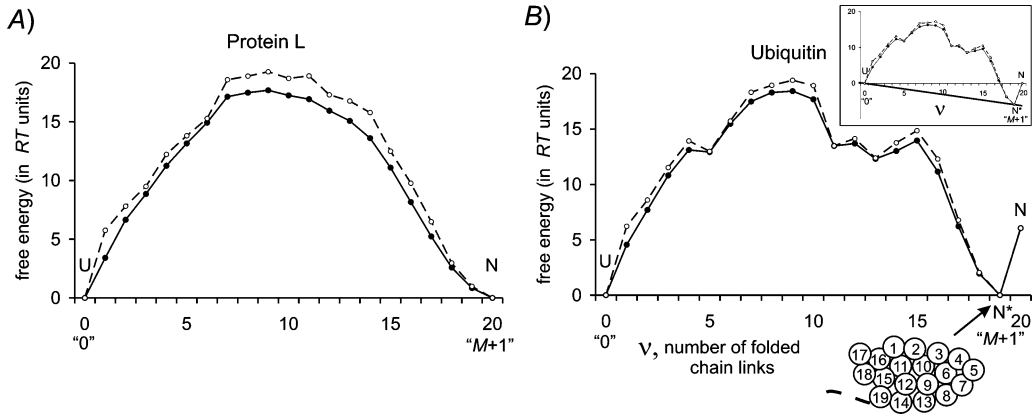

Figure 2. Simplified two-dimensional representations of the free energy barriers for folding−unfolding of protein L (A) and ubiquitin (B). In both cases, the polypeptide chain is divided into 20 links (a "link" here has 3 residues for protein L (consisting of $L = 60$ residues) and 3−4 residues for ubiquitin (consisting of $L = 76$ residues)). The lowest free-energy value for a microstate having a given number of folded links is presented by open circles and a dashed line; the free energy for the ensemble of all microstates having the given number of folded links is presented by solid circles (where the open circles do not shadow them) and a black line. The plots are drawn for the case when the two most stable and separated by the free energy barrier microstates of the protein chain have equal free energies. For protein L, these are the completely unfolded state U and the completely folded state N; in this case, $(\varepsilon/RT)_{mt} = (\varepsilon/RT)_0$ (see eq 37). For ubiquitin, these are the completely unfolded U and N*, one of the nearly folded states; in this case, $(\varepsilon/RT)_{mt}$, which slightly differs from $(\varepsilon/RT)_0$ (see eq 38), is used to obtain equal free energies for U and N*. The ubiquitin plot obtained with $(\varepsilon/RT)_0$ determined by eq 37 is shown in the inset; here the tangent connects the two deepest free energy minima, U and N*. The nearly folded stable microstate N* of ubiquitin (with 19 folded chain links) is schematically drawn under the horizontal axis. The lowest-free-energy unfolded (or nearly unfolded) and folded microstates (U and N for protein L; U and N* for ubiquitin) are numbered as "0" and "$M + 1$", respectively.

taken as l_0 plus a rounded integer of $n \times (L/N)$, and then $l'_{n+1} = l''_n + 1$ is the chain number of the first residue of the next link $n + 1$.

In all cases when a protein chain includes $L \geq 20$ amino acid residues, the number of links $N = 20$; when $L < 20$, $N = L$.

Computational Limitations of the Model. We can use chains with no more than $N \approx 20$ links in our computations, or rather, we cannot use networks including more than $\sim 2^{20}$ microstates, because solution of a set of more than $\sim 2^{20}$ equations takes too long. On the other hand, "elementary links" used in considerations underlying the presented theory must be more or less compact. Otherwise, all or the majority of semifolded intermediates forming a network of folding−unfolding pathways would be very much eroded rather than compact, and therefore, their energy would be too high due to a simple artifact, that is, due to too large and noncompact links. Thus, the chain length of a link should not exceed the diameter of a globule. The chain length of one link is $x = l_1(L/N)$, where l_1 is the chain length of one residue, while the diameter of a globule is $\sim(LV_1)^{1/3}$, V_1 being the volume of an average amino acid residue. The limitation $x < (LV_1)^{1/3}$ means that $L < (N^3V_1/l_1^3)^{1/2}$. Using the typical values[31] $V_1 = 150$ Å^3, $l_1 = 3.8$ Å, and $N = 20$ links-per-protein limit, we see that we hardly can accurately simulate transitions in proteins of $L \approx 150$ residues or more. For long protein chains, we will probably overestimate free energy barriers and therefore underestimate the folding rates.

Determination of the Midtransition Point. This paper deals with the simplest (cf. refs 7 and 31) case of protein folding and unfolding at midtransition, where the rates of these two processes are equal. Here, at the point of thermodynamic equilibrium between the folded and unfolded states, free energies of the starting (unfolded or nearly unfolded) microstate "0" and the final (folded or nearly folded) microstate "$M + 1$" must coincide: $G_0 = G_{M+1}$.

In the simplest case, folding starts from the free energy minimum corresponding to the completely unfolded microstate U (with $E_U = 0$, $S_U = sL$) and finishes at the minimum corresponding to the completely folded microstate N (with $E_N = \varepsilon\nu_N$, $S_N = s\cdot(L - m_N) + \sum_{\text{unresolved loops}} S_{\text{loop}}$, where the last sum is taken over internal "loops" formed by $L - m_N$ residues with unresolved coordinates within the native structure; S_{loop} is computed after eq 35). Then $(\varepsilon/RT)_{mt}$, the midtransition value of ε/RT (note that all our calculations use dimensionless free energies $G_i/RT = E_i/RT - S_i/R$) is determined as $(\varepsilon/RT)_{mt} = (\varepsilon/RT)_0$, where

$$(\varepsilon/RT)_0\nu_N = S_N/R - S_U/R \tag{37}$$

with ν_N reflecting the number of contacts in a completely folded protein chain (while a completely unfolded protein chain has no contacts), and $S_U = sL$ (while $S_N = s\cdot(L - m_N) + \sum_{\text{unresolved loops}} S_{\text{loop}}$).

Having an "all-or-none" native state-to-coil transition, we should expect that the two most stable microstates correspond to the native state N and the unfolded state U (Figures 1A and 2A). However, given crude free energy estimates that we use, we may observe that, sometimes, an "almost completely unfolded" microstate U* is estimated as being more stable than the completely unfolded state U, and/or an "almost completely folded" microstate N* is estimated as being more stable than the completely folded state N (Figure 2B; see also inset). Since our main math tool, a quasi-stationary approximation, requires considering transition between the two most stable microstates, we have, in these cases, to ascribe indexes "$M + 1$" and "0" not to N and U microstates, but to the lowest-free-energy N* and U* microstates, respectively. Further, to make "$M + 1$" and "0" microstates have equal free energies, we have to modify the $(\varepsilon/RT)_0$ value using an evident generalization:

$$(\varepsilon/RT)_{mt}[\nu_{M+1} - \nu_0] = S_{M+1}/R - S_0/R \tag{38}$$

This modification can proceed in a few iterations, if necessary, until a thermodynamic equilibrium between the two lowest-

free-energy minima, "$M + 1$" and "0", is achieved. As a result, the final value of $(\varepsilon/RT)_{mt}$ is obtained.

The folding rate, etc., calculation can be started, if the obtained microstates "$M + 1$" and "0" are separated by a free energy barrier and unconnected by an elementary transition. Otherwise, we have to conclude that the used free energy estimates contradict to existence of an "all-or-none" transition in the given protein, and the developed algorithm with given free energy estimates cannot be applied to computation of its folding rate and nucleus.

It should be noted that the $(\varepsilon/RT)_{mt}$ values are protein-specific. It comes as no surprise, since different proteins have different amino acid residue contacts and therefore different average strengths of atom−atom interactions described by ε.

It should also be noted that a deviation of ε/RT from $(\varepsilon/RT)_{mt}$ for a given protein corresponds to a deviation from the midtransition conditions. Theoretical investigation of protein folding and unfolding that occur far from the midtransition we postpone until the next paper of this series.

Materials

Our study includes all single-domain proteins, separate domains of large proteins and two short polypeptides (α-helix and β-hairpin) with experimentally studied folding kinetics, which (i) are monomeric, (ii) do not have disulfide bonds within the protein chain or covalent bonds with ligands, and (iii) have known 3D structures. Choice of the used PDB entries is done as follows: for proteins with extensively studied folding nuclei (they are listed in Figure 4, see below), PDB entries are taken from Garbuzynskiy et al.;[20] otherwise, we take PDB entries from ref 3, which contains recommended PDB entries for many two-state proteins, or from our previous studies.[32,33] In two cases, when there was no exact PDB entry correspondence to the chains with experimentally studied folding process, we used PDB structure of their closest analogs (see footnotes (a), (b) to Table S1 of Supporting Information). All these proteins with their folding−unfolding rates in water, folding rates at midtransition, and other kinetic data are collected in the database "KineticDB".[34]

The experimental values for the midtransition folding (=unfolding) rates, k_f^{mt}, are taken from the original papers quoted in Table S1, Supporting Information. For proteins with multistate kinetics, we consider the rate constants that belong to the slowest phase of transitions (but for those attributed solely to cis−trans proline isomerization).

The ϕ values for proteins with extensively studied folding nuclei (their numbers are underlined in Table S1, Supporting Information) are taken from ref 20. In most cases these values have been determined for "in-water" conditions rather than for the midtransitions.

Results and Discussion

In this paper we consider the in vitro protein folding rates and nuclei for single-domain globular proteins under midtransition conditions.

Experimentally, many small and middle-sized proteins exhibit in vitro an "all-or-none" or "two-state" transition[24,27,31] between their native and denatured states; that is, only these two macroscopic states of these proteins are stable and present a visible amount at midtransition, where, after the kinetic processes are over, they are in equal quantities. However, with only one protein state present and the other absent at the beginning, the rate of approaching to equilibrium can be determined. Moreover, for "multi-state" proteins,[23,24] all metastable intermediates (that can complicate the folding analysis) are usually unstable at midtransition and, therefore, virtually do not show up during the midtransition folding (or unfolding), even if they accumulate and can be observed far from the midtransition, when the native state becomes much more stable than the unfolded one.[4] This is a great advantage of the midtransition folding for theoretical studies.[7,8,11,20] Thus, this paper considers the basic case (folding/unfolding at midtransition), while the next paper of the series will consider protein folding that occurs far from midtransition, which is more physiologically relevant.

Checking the Existence of an "All-or-None" Transition with the Used Free Energy Estimates. To model the midtransition conditions for a given protein, we have to choose (see eqs 37 and 38) an ε/RT value that would provide two free energy minima for this protein chain ("$M + 1$", corresponding to the folded or nearly folded state, and "0", corresponding to the unfolded or nearly unfolded state), with $G_{M+1} = G_0$ (see Figures 1 and 2).

Before other calculations, we have to know whether or not the landscape of folding−unfolding pathways with the used free energy estimates contains these two minima and a barrier between them. For this purpose, the natively folded part of the chain of each protein was divided into "chain links", as described under "Chain division into "links"". Further, for each protein, we calculated the $(\varepsilon/RT)_{mt}$ values corresponding to midtransitions between its two most stable microstates (see "Determination of the Mid-Transition Point"). If completely folded or/and completely unfolded microstates did not correspond to the deepest free-energy minima, we singled out the most stable nearly folded and nearly unfolded microstates, as described.

The analysis has failed to reveal any protein (out of 82 studied) where the microstates "0" and "$M + 1''$", corresponding to the two deepest free-energy minima are not separated by a free energy barrier (see Figure 2A,B as examples). The situation shown in Figure 2A is true for 44 proteins (## 3, 5, 12, 16−20, 23, 25, 26, 32, 35, 36, 38, 40, 44−49, 51, 54, 56−64, 66−70, 74, 76, 77−79, 81 in Table S1, Supporting Information), while the situation shown in Figure 2B occurs in 38 cases (## 1, 2, 4, 6−11, 13−15, 21, 22, 24, 27−31, 33, 34, 37, 39, 41−43, 50, 52, 53, 55, 65, 71−73, 75, 80, 82).

Folding Rates at Midtransition. Thus, our calculations described under "Algorithm for Calculation of the Folding and Unfolding Rates" and "Algorithm for Calculation of the Folding Nuclei" are applicable to all 82 studied proteins, and we can calculate the midtransition folding rates k_f^{mt} $(=k_u^{mt})$ of these proteins using eq 9.

A comparison of logarithms of the calculated midtransition folding rates with experiment is given in Figure 3. Since our method is developed, actually, for two-state protein folding, we compare the calculated and experimentally measured folding rates for two- and multistate proteins separately (see Figure 3A,B, respectively). The obtained agreement with experiment is better for two-state proteins than for multistate ones. The theory-to-experiment correlation coefficient is 0.81 for the two-state folders, and only 0.56 for the three-state folders (and 0.79 for the totality of examined proteins; see Figure 3C). Possibly, the worse prediction of the folding rates for multistate proteins partially resulted from the fact that these proteins have metastable intermediates enhanced by non-native interactions, which we ignored in our theory. However, the data presented in Figure 3 suggest that the main reason for the worse prediction is merely the larger size of these proteins (see "Computational Limitations of the Model"). Indeed, after exclusion of proteins with $L > 150$ residues, the correlation becomes 0.85 for two-

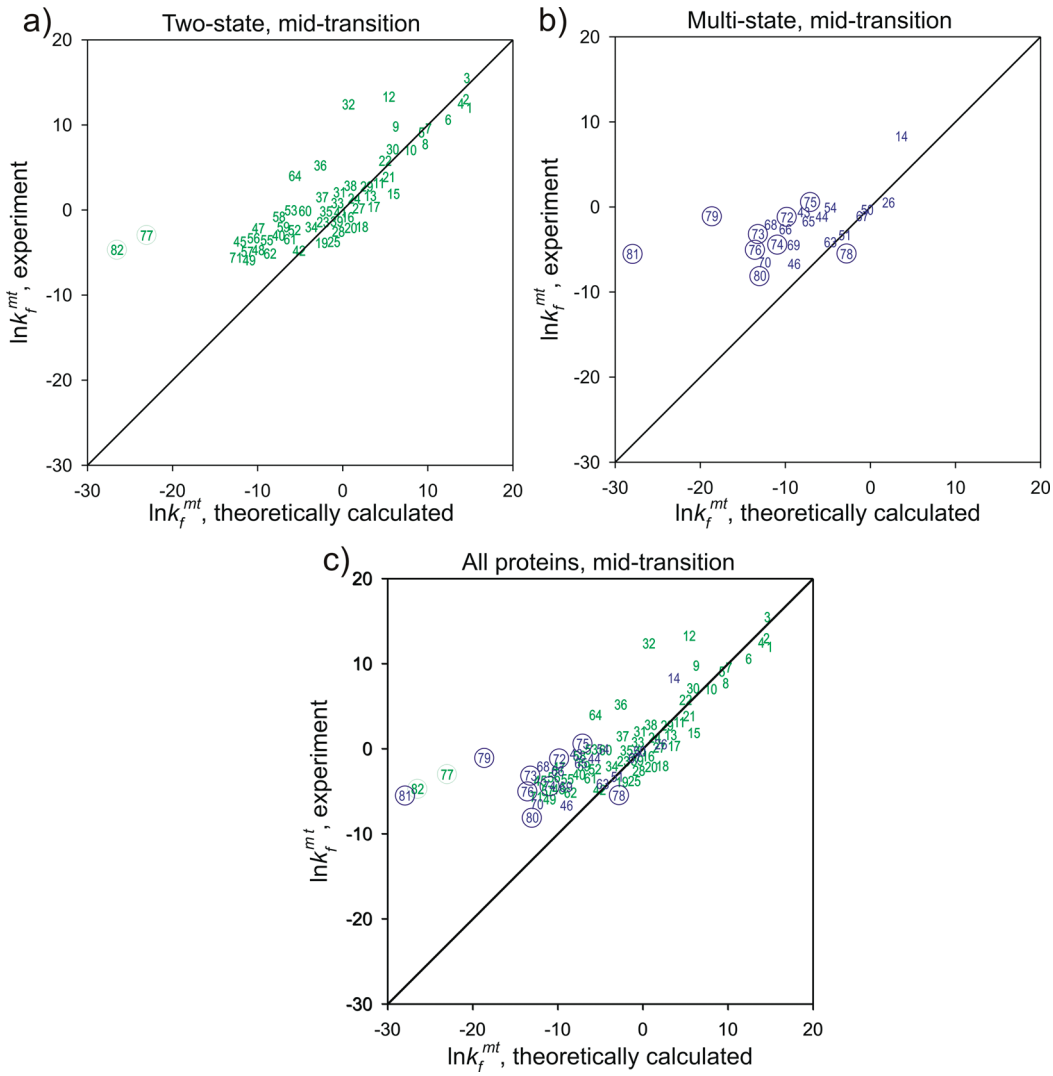

Figure 3. Comparison of calculated (horizontal axis) and experimentally measured (vertical axis) logarithms of the folding rates in midtransition (mt) for two-state proteins (A), multistate proteins (B), and the totality of all proteins (C). Proteins are drawn by numbers # used in Table S1, Supporting Information. The correlation coefficient is 0.81 ± 0.05 for two-state proteins (0.78 ± 0.05 when the α-helix and the β-hairpin, which are merely the secondary structure elements, are excluded), 0.56 ± 0.07 for multistate proteins, and 0.79 ± 0.04 for the totality of all proteins and short peptides. After exclusion of proteins with >150 folded residues (encircled, see the text), the correlation becomes 0.85 ± 0.04 for two-state proteins (0.83 ± 0.04 without the α-helix and the β-hairpin), 0.73 ± 0.07 for multistate proteins, and 0.85 ± 0.03 for all proteins and peptides.

state proteins, 0.73 for multistate proteins, and 0.85 for all such proteins and peptides.

One can see that the calculated folding rates are much lower than the experimental ones for slow-folding proteins that are usually large. We think that theoretical underestimation of the folding rates of large proteins results from the fact that the size of one "chain link" used in our computations exceeded the diameter of a globule of proteins with more than 150 residues (see "Computational Limitations of the Model"), and hence, the chain links could not be considered as compact "beads" implicated by our theory in this case. Indeed, any semifolded globule formed by too large "links" becomes rather eroded, which is an artifact. If we exclude all proteins of more than 150 residues (encircled in Figure 3), we obtain a little better theory-to-experiment correlation for two-state folders, and significantly better correlation for three-state folders many of which are large (see Figure 3).

It should be stressed that, apart from 3D protein structures, our calculations contain only two important experimental parameters, $s = 2.3R$,[27] and $k^* = 10^8$ s^{-1},[30] and do not contain any adjustable parameters at all, and, nevertheless, most of the points in the plots (especially for proteins of ≤150 residues) are grouped around diagonals of these plots (i.e., around the lines where the theoretically calculated and experimental folding rates not only correlate but coincide).

Computation of Folding Nuclei: ϕ Values at Midtransition. To compare the calculated and experimental results on folding nuclei, we took 17 proteins reported by ref 20, whose folding nuclei are well studied experimentally using point mutations (proteins ## 16, 17, 19, 23−25, 44−46, 50, 53−55, 62, 63, 65, 67 underlined in Table S1, Supporting Information). As usual,[20,23] we take into consideration only those mutations, which reduce the size of side chains. Experimental ϕ values for these mutations are listed in Table 1 of ref 20.

The theoretical ϕ values have been computed from eq 29 using PDB structures of 17 proteins underlined in Table S1, Supporting Information. For each protein, the mutation-independent values P_i, used in eqs 29, 31, were calculated for the "wild type" form, while the mutation-dependent values $[\delta\nu_i]_\alpha/[\delta\nu_{M+1}]_\alpha$, used in eqs 29, 30, were calculated for each separate mutant listed in Table 1 of ref 20.

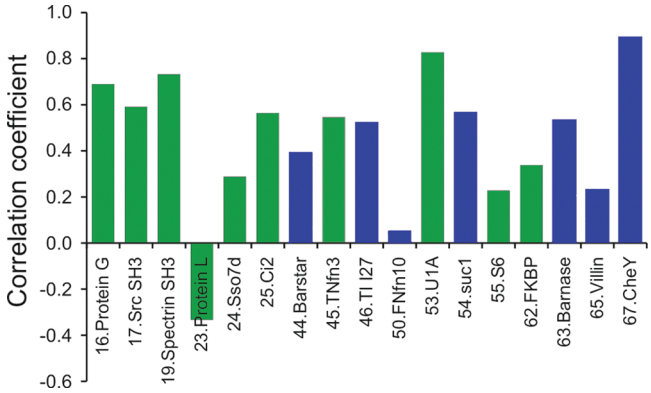

Figure 4. Correlation between the calculated (at midtransition) and experimentally measured ϕ values for 17 proteins with experimentally extensively explored folding nuclei (see ref 20). Green bars correspond to two-state proteins; blue bars correspond to multistate proteins. The average correlation amounts to 0.45 ± 0.29 (0.45 ± 0.33 for two-state folders, 0.46 ± 0.26 for multistate folders).

In this work we calculated ϕ values at midtransition, but we had to compare them with the reported experimental ϕ values that, for most proteins, were extrapolated to "in-water" (## 19, 24, 25, 44, 50, 53, 54, 63, 67) or nearly "in-water" (## 16, 17, 23, 45, 46, 55) conditions. ϕ values of only two proteins (## 62, 65) corresponded to midtransition conditions. A calculation of the "in-water" ϕ values will be presented in the next paper of this series.

The results of comparison with experiment are given in Figure 4. One can see that the quality of prediction changes dramatically from protein to protein. On the average, the correlation coefficient is 0.45, which is not bad, but still a little worse than in the previous work[20] where we limited intermediate microstates to the ones containing two or less internal loops only, which allowed us to use smaller links.

The fact that the calculated ϕ values correlate with experiment worse than the folding rates is a result of ϕ values ranging from 0 to 1, while the range of the folding rates covers 10 orders of magnitude. Actually, in a situation of two parallel folding pathways, where one pathway crosses "the main" transition state, which is more stable than the other by one *RT* unit only, erroneous destabilization of the more stable transition state by a couple of *RT* only (which, in calculations, can be easily done by our approximate energy estimates) erroneously switches protein folding to another pathway and dramatically changes ϕ values describing the folding nucleus. Yet, in this case, the folding rate changes by a factor of $\sim$3 only, which is almost nothing against the observed billion-time variety of folding rates. As follows from this example, even small errors in the free energy estimates can lead to a wrong ϕ value prediction for proteins with multiple pathways, while prediction of the folding rates can remain successful enough.

The above reasons are applicable not only to errors in calculations but also to perturbations caused by changes in the protein chain sequence. As an example, let us consider protein G and protein L. These proteins have very similar 3D structures, their folding rates differ by $\approx$10 times only, but their folding nuclei show different localization within the 3D structure.[35,36] Prediction of the folding rates is satisfactory for both protein G (#16) and protein L (#23); see Figure 3A. On the contrary, prediction of ϕ values, being satisfactory for protein G, is extremely bad for protein L (see Figure 4), where the nucleus is predicted to be found almost in the same place as in protein G. In this connection, it is pertinent to note that the folding nucleus of protein G can be switched by mutations to the position of the folding nucleus of protein L.[37]

Conclusion

The calculation-to-experiment correlations obtained in this work are at least not worse than those achieved in phenomenological theories by other groups,[38−40] but the ability to estimate the absolute values of the folding rates using no adjustable parameters and, simultaneously, to outline the folding nuclei is a unique feature of works based on the landscape theory of protein folding[7] (see also refs 8, 11, 12, and 20).

In this study we have extended and generalized the analysis of in vitro folding and unfolding flows[8] to calculate the protein folding rates and to outline the folding nucleus. The flow model extends the conventional, nucleation-based description of protein folding. The proposed model has no adjustable parameters at all. Rather, it exploits only 3D protein structures and the experimentally found rate of an elementary conformational rearrangement and entropy loss at folding of one amino acid residue. Nevertheless, and despite the roughness of the free energy estimates, the correlation between the calculated and experimentally measured protein folding rates and ϕ values is reasonably high: we have an average correlation of 0.79 (0.85 after exclusion of proteins with $L > 150$ residues) for prediction of the protein folding rates and a correlation of 0.45 for prediction of the protein folding nuclei via ϕ values.

Acknowledgment. We thank V. V. Filimonov, E. Paci, O. V. Galzitskaya, and S. O. Garbuzinskiy for helpful discussions and E. V. Serebrova for assistance in manuscript preparation. This work was supported by grant of the President of Russian Federation (MK-4894.2009.4), by programs "Molecular and Cellular Biology" of the Russian Academy of Sciences and "Leading Scientific Schools of Russia" (grant NSh-2791.2008.4), by a grant from the Federal Agency for Science and Innovation (02.740.11.0295), by Russian Foundation for Basic Research (grant 07-04-00388), by the INTAS (grant 05-1000004-7747) and by an International Research Scholar's Award 55005607 to A.V.F. from the Howard Hughes Medical Institute.

Supporting Information Available: List of studied proteins with their experimental and calculated folding rates and other relevant details. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952.
(2) Goldberg, M. E.; Semisotnov, G. V.; Friguet, B.; Kuwajima, K.; Ptitsyn, O. B.; Sugai, S. *FEBS Lett.* **1990**, *263*, 51.
(3) 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.; Vallee-Belisle, 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.; Plaxco, K. W. *Protein Sci.* **2005**, *14*, 602.
(4) Jackson, S. E. *Folding Des.* **1998**, *3*, R81.
(5) Galzitskaya, O. V.; Ivankov, D. N.; Finkelstein, A. V. *FEBS Lett.* **2001**, *489*, 113.
(6) Munoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311.
(7) Finkelstein, A. V.; Badretdinov, A. *Folding Des.* **1997**, *2*, 115.
(8) Ivankov, D. N.; Finkelstein, A. V. *Biochemistry* **2001**, *40*, 9957.
(9) Alm, E.; Morozov, A. V.; Kortemme, T.; Baker, D. *J. Mol. Biol.* **2002**, *322*, 463.
(10) Alm, E.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11305.

(11) Galzitskaya, O. V.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11299.
(12) Galzitskaya, O. V.; Skoogarev, A. V.; Ivankov, D. N.; Finkelstein, A. V. *Pac. Symp. Biocomput.* **2000**, 131.
(13) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Finkelstein, A. V. *J. Phys. Condens. Matter* **2005**, *17*, S1539.
(14) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.
(15) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258.
(16) Lei, H.; Duan, Y. *J. Mol. Biol.* **2007**, *370*, 196.
(17) Lei, H.; Duan, Y. *J. Phys. Chem. B* **2007**, *111*, 5458.
(18) Day, R.; Daggett, V. *J. Mol. Biol.* **2007**, *366*, 677.
(19) Schaeffer, R. D.; Fersht, A.; Daggett, V. *Curr. Opin. Struct. Biol.* **2008**, *18*, 4.
(20) Garbuzynskiy, S. O.; Finkelstein, A. V.; Galzitskaya, O. V. *J. Mol. Biol.* **2004**, *336*, 509.
(21) Emanuel, N. M.; Knorre, D. G. *Course of Chemical Kinetics*; Vyschaya shkola: Moscow, 1984.
(22) Bahvalov, N. S.; Zhidkov, N. P.; Kobel'kov, G. M. *Numerical Methods*; Nauka: Moscow, 1987.
(23) Matouschek, A.; Kellis, J. T., Jr.; Serrano, L.; Bycroft, M.; Fersht, A. R. *Nature* **1990**, *346*, 440.
(24) Fersht, A. R. *Curr. Opin. Struct. Biol.* **1995**, *5*, 79.
(25) Taketomi, H.; Ueda, Y.; Go, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445.
(26) Krieger, E.; Koraimann, G.; Vriend, G. *Proteins* **2002**, *47*, 393.
(27) Privalov, P. L. *Adv. Protein Chem.* **1979**, *33*, 167.
(28) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Interscience: New York, 1969.
(29) Metropolis, N.; Rosenbluth, M. N.; Rosenbluth, A. W.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.
(30) Zana, R. *Biopolymers* **1975**, *14*, 2425.
(31) Finkelstein, A. V.; Ptitsyn, O. B. *Protein Physics*; Academic Press: Amsterdam - Boston - London - New York - Oxford - Paris - San Diego - San Francisco -Singapore - Sydney - Tokyo, 2002.
(32) Ivankov, D. N.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 8942.
(33) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. *Protein Sci.* **2003**, *12*, 2057.
(34) Bogatyreva, N. S.; Osypov, A. A.; Ivankov, D. N. *Nucleic Acids Res.* **2009**, *37*, D342.
(35) Kim, D. E.; Fisher, C.; Baker, D. *J. Mol. Biol.* **2000**, *298*, 971.
(36) McCallister, E. L.; Alm, E.; Baker, D. *Nat. Struct. Biol.* **2000**, *7*, 669.
(37) Nauli, S.; Kuhlman, B.; Baker, D. *Nat. Struct. Biol.* **2001**, *8*, 602.
(38) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985.
(39) Gromiha, M. M.; Selvaraj, S. *J. Mol. Biol.* **2001**, *310*, 27.
(40) Zhou, H.; Zhou, Y. *Biophys. J.* **2002**, *82*, 458.

JP912186Z