PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Analyzing the effect of homogeneous frustration in protein folding

Vinícius G. Contessoto,[1†] Debora T. Lima,[1†] Ronaldo J. Oliveira,[1,2] Aline T. Bruni,[3] Jorge Chahine,[1] and Vitor B. P. Leite[1*]

[1] Departamento de Física, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, Sao José do Rio Preto, São Paulo 15054-000, Brazil

[2] Laboratório Nacional de Ciência e Tecnologia do Bioetanol, Campinas, São Paulo 13083-970, Brazil

[3] Departamento de Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, São Paulo, Brazil

ABSTRACT

The energy landscape theory has been an invaluable theoretical framework in the understanding of biological processes such as protein folding, oligomerization, and functional transitions. According to the theory, the energy landscape of protein folding is funneled toward the native state, a conformational state that is consistent with the principle of minimal frustration. It has been accepted that real proteins are selected through natural evolution, satisfying the minimum frustration criterion. However, there is evidence that a low degree of frustration accelerates folding. We examined the interplay between topological and energetic protein frustration. We employed a C_α structure-based model for simulations with a controlled nonspecific energetic frustration added to the potential energy function. Thermodynamics and kinetics of a group of 19 proteins are completely characterized as a function of increasing level of energetic frustration. We observed two well-separated groups of proteins: one group where a little frustration enhances folding rates to an optimal value and another where any energetic frustration slows down folding. Protein energetic frustration regimes and their mechanisms are explained by the role of non-native contact interactions in different folding scenarios. These findings strongly correlate with the protein free-energy folding barrier and the absolute contact order parameters. These computational results are corroborated by principal component analysis and partial least square techniques. One simple theoretical model is proposed as a useful tool for experimentalists to predict the limits of improvements in real proteins.

Proteins 2013; 81:1727–1737.

Key words: structure-based model; multivariate analysis; molecular dynamics; C-alpha model.

INTRODUCTION

The protein folding problem is an important challenge in the understanding of biomolecular mechanisms. In the late 1960s, Levinthal[1] raised the question of how many configurational states a typical protein accesses before it dynamically reaches its lowest energy-folded state. In the last decades, based on the energy landscape theory and on the concept of folding funnels, the apparent Levinthal folding paradox has been solved.[2–6] The energy landscape has been revealed to be a robust framework in the qualitative and quantitative protein-folding studies in theoretical[7–9] as well as experimental[10–13] works. In addition, protein folding has been investigated using different computational models to predict rates and stability to correlate with experimentalists.[8,14,15]

According to the theory, protein folding is described as the diffusion of an ensemble of partially folded structures through which a protein passes on its way to the native state.[16–18] The overall landscape resembles a funnel with some roughness, reflecting transient traps in local energy minima.[19] The folded native state is associated with the global energy minimum of the system.[20] To be kinetically foldable, the funnel must have a steep enough slope to be able to overcome local traps, leading

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

[†] Vinicius G. Contessoto and Debora T. Lima contributed equally to this work.

*Correspondence to: Departamento de Física Instituto de Biociências, Letras e Ciências Exatas Universidade Estadual Paulista Sao Jose do Rio Preto, São Paulo 15054-000, Brazil. E-mail:vleite@sjrp.unesp.br

Received 26 September 2012; Revised 13 March 2013; Accepted 18 March 2013
Published online 23 April 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24309

to the global minimum.[21] In spin glass theory, as well as in other scientific areas, frustration occurs owing to the impossibility of satisfying all favorable energetic interactions simultaneously and it is relevant for protein folding.[22–24] Biological proteins, selected through the process of natural evolution, seek to maximize favorable energetic interactions of the native configuration, following the principle of minimal frustration.[16,22–24] Consequently, naturally occurring proteins are considered minimally frustrated systems. However, frustration derived from unfavorable interactions should not be neglected. Frustration can lead to traps and to intermediate states.[25] On the other hand, the presence of frustration influences the folding dynamic and can be favorable in accelerating its kinetics.[26–30] Another aspect of the theory is that local frustration can be a useful tool owing to the fact that frustrated regions on the protein surface are related to substrate-binding sites.[31,32] In principle, the existence of frustration hinders folding kinetics and stability. However, studies show that a low degree of frustration helps the protein-folding process.[27,33–35]

Frustration can be divided into two types: topological and energetic. Topological frustration is associated with the structure, it is intrinsic to each protein, and depends on each fold or motif.[36] This aspect is hard to control in protein computational models. On the other hand, energetic frustration is associated with the energy interaction between monomers and can be simulated by adding interaction terms to the potential energy function.[7,27–29,34,37,38]

In this study, we explore the effects of frustration in kinetics and thermodynamics using structure-based model proteins different in size and structural motifs. The main goal is to determine the conditions under which the presence of frustration is favorable to the folding process. We determined the criteria that distinguish these conditions and our findings correlate with the concept of absolute contact order,[39] free-energy folding barrier,[32] and folding time.[8,14] The theoretical framework developed here should be a valuable tool to help experimentalists to distinguish proteins of which thermal stability and rates of folding could be enhanced.

METHODS

Structure-based C_α model

In this study, proteins are coarse-grained in C_α atom level of simplification.[9,40] Protein residues are represented as single beads located at the C_α atom position in the protein main chain. The Hamiltonian that gives the protein interaction energy is based on the geometry of its native state, such that the potential energy surface reaches its minimum at this reference state.[41] Thus, the model has no intrinsic energetic frustration. The functional form of the potential of a given structure Γ with respect to its native structure Γ_o is defined as

$$V(\Gamma,\Gamma_o)=\sum_{\text{bonds}}\varepsilon_r(r-r_o)^2+\sum_{\text{angles}}\varepsilon_\theta(\theta-\theta_o)^2 + \sum_{\text{dihedrals}}\varepsilon_\varphi\left\{[1-cos(\varphi-\varphi_o)]+\frac{1}{2}[1-cos(3(\varphi-\varphi_o))]\right\} + \sum_{\text{contacts}}\varepsilon_C\left[5\left(\frac{d_{ij}}{r_{ij}}\right)^{12}-6\left(\frac{d_{ij}}{r_{ij}}\right)^{10}\right]+\sum_{\text{non-contacts}}\varepsilon_{NC}\left(\frac{\sigma_{NC}}{r_{ij}}\right)^{12} \quad (1)$$

with $\varepsilon_r = 100\varepsilon_0$, $\varepsilon_\theta = 20\varepsilon_0$, $\varepsilon_\phi = \varepsilon_0$, $\varepsilon_C = \varepsilon_0$, $\varepsilon_{NC} = \varepsilon_0$, $\sigma_{NC} = 4.0$ Å, and ε_0 is the interaction energy per contact, which is 1 unit (in reduced units) by construction. r_o, θ_o, ϕ_o, and d_{ij} are values extracted from the native coordinate structure. In Eq. ((1)), adjacent beads and bond angles interact via harmonic terms, first and second terms, respectively. The third term represents the dihedral rotation of the chain. r_{ij} is the geometric distance between two beads. The interaction of the nonbonded residues pairs in contact in the native state is given by the Lennard–Jones 10–12 potential. All residue pairs which are not in contact in the native structure interact via nonspecific repulsion. d_{ij} is the distance for two residues i and j in the native configuration with the native contact map determined by the software contact of structural units (CSU).[42] Structure-based models have been successful in predicting folding mechanisms such as experimental φ-values[10,43] and other experimental quantities in agreement with theory.[40] Simulations employing structure-based models show a strong correlation of the calculated folding rates with free-energy barriers.[14,15] Simulations have also shown how important it is to consider the diffusion coefficient as being dependent on the reaction coordinate to improve the accuracy of the folding rate prediction.[18,34,44–47] Folding rates have excellent correlation with dimensionless parameters related to protein topology, such as the contact order.[8,39]

Frustration potential term

Structure-based models are free of nonspecific energetic frustration as all interactions used to construct the potential model lead the protein to its native state, the potential energy minimum.[26,27,32,34] To include nonspecific energetic interactions, an extra potential term [Eq. (2)] is added to the potential [Eq. (1)].[28,29,34,37] This term has an attractive interaction between all residue pairs not in contact in the native configuration. Thus, homogeneous energetic frustration between monomers at a certain distance is introduced. The potential form of the frustration interaction between monomers i and j is Gaussian-like, given by

$$V_f(r_{ij})=-\varepsilon_f\exp\left\{-\frac{(r_{ij}-\bar{d})^2}{\sigma_f^2}\right\}, \quad (2)$$

Table I
Data of 19 Proteins Simulated and Sorted by the Product $\Delta F \times$ ACO

Name	PDB[a]	# a.a.	Q[b]	$\bar{d}$ (Å)[c]	ACO[d]	$\Delta F(k_B T_f)$[e]	$\Delta F \times$ ACO	$\varepsilon_f^{\text{opt}}$[f]
ACR	1ARR	53	57	7.240	2.66	0.00	0.00	0.00
HP36	1VII	36	56	7.180	3.97	0.00	0.00	0.00
PSBD	2PDD	43	62	7.080	4.75	0.00	0.00	0.00
$\alpha_3 D$	2A3D	73	136	7.690	6.77	0.20	1.35	0.00
PtABD	1BDC	60	102	7.670	5.20	0.35	1.82	0.00
EnHD	1ENH	54	111	8.081	6.80	0.75	5.10	0.00
IM9	1IMP	86	174	7.330	9.80	1.70	16.66	0.05
ACBD	2ABD	86	182	7.087	11.76	1.50	17.64	0.05
HHCC	1HRC	104	244	7.350	11.59	1.90	22.02	0.05
PtL	2PTL	60	134	6.957	10.87	2.50	27.17	0.10
PtG	2K0P	56	133	7.026	9.56	3.50	33.46	0.10
ADA2h	1PBA	81	172	7.400	11.85	3.00	35.55	0.10
CI2	1CIS	66	151	7.060	10.71	3.75	40.16	0.10
SH3	1FMK	61	148	7.078	11.30	3.90	44.07	0.10
Ubiquitin	1UBQ	76	182	7.318	11.40	4.00	45.60	0.05
CSPTm	1G6P	66	166	6.475	11.68	4.90	57.23	0.20
HPr	1HDN	85	214	6.834	15.61	4.90	76.49	0.15
αAIT	2AIT	74	182	6.593	14.33	6.00	85.98	0.25
TWIg	1WIU	93	220	6.530	18.78	5.20	97.65	0.25

[a] Protein Data Bank ID.
[b] Total number of native contacts.
[c] Average over contact distances for the native structure (PDB coordinates).
[d] Absolute contact order.[39]
[e] Free-energy barrier between unfolded and folded valleys on the free-energy profile as a function of Q.
[f] The optimal value for the energetic frustration parameter [Eq. (2)].

and centered at the average contact distance between monomers in the native configuration $\bar{d}$ (for the particular value of each protein, see Table I). In Eq. (2), r_{ij} accounts for distances between any pair of monomers except the native contact ones. $\sigma_f = 1.0$ Å by definition. The intensity of the frustration potential term is determined by ε_f in units of energy per contact, ε_0. ε_f was studied in the interval of $0.0 \leq \varepsilon_f \leq 0.5$ for the set of proteins in this study. Similar methodologies, with modified versions of C_α structure-based model, were employed to study non-native interactions using non-native hydrophobic interactions[28] and non-native electrostatic interactions.[29,38] Frustration occurs when nonspecific interactions compete with native interactions, thus increasing the roughness of the energy landscape.[36] One could argue in favor of other types of energetic frustration potentials. However, this simple potential has an appealing feature, which is its homogeneity, so that the results do not need to be averaged over different initial potential configurations. Moreover, as we will show, our main qualitative result is not strongly dependent on the type of energetic frustration we employ.

Absolute contact order

Introduced by Baker and coworkers,[39] the absolute contact order is determined by the sum of the sequence distance between all residues in contact divided by the total number of formed contacts

$$\text{ACO} = \frac{1}{N_c} \sum_{(ij)}^{N_c} |i-j|, \tag{3}$$

with N_c being the number of contacts and $|i-j|$ the sequence separation distance between residues i and j in contact. Absolute contact order (ACO) gives an idea of the importance of local and nonlocal interactions for the protein structure and correlates with entropic constraint created as contacts are formed. ACO normalized by the number of the protein residues is the relative contact order. Contact order is easily and quickly calculated by the algorithm available online.[48] For the contact order calculation, only the protein coordinates in the Protein Data Bank format are required.[49] It is also important to characterize the free-energy barrier height for folding which has strong correlation with folding rates.[39,50]

Simulation details

Structure-based model input files are obtained using the structure-based models in gromacs (SMOG) webtool[51] and simulations are performed using the molecular dynamic package GROMACS.[52] Proteins are initialized in an open random configuration and simulated over 5×10^8 steps with time steps equal to 0.5 fs and equilibrated after 1×10^7 steps. Configurations are stored every 1000 steps. The Berendsen thermostat algorithm[53] is employed to maintain coupling to an external bath with constant equal to 1 ps. The thermodynamic free-energy profile is obtained by combining multiple simulations performed over a range of constant temperature runs using the weighted histogram analysis method (WHAM).[54] The reaction coordinate used to monitor folding events is defined as the sum of native contacts (Q) on a given structure Γ at time t (see an example of Q as a function of time in Supporting Information Fig. S1) One native contact between two residues i and j is considered formed when the distance between them is shorter than $1.2d_{ij}$. It has been shown that Q is a reliable reaction coordinate for probing small, single-domain protein-folding dynamics.[55] The number of topological contacts (C_t) is also used as a reaction coordinate. One topological contact is formed if the distance between two residues is in the region where the frustration potential acts [Eq. (2)]. Thus, even for $\varepsilon_f = 0.0$ in Eq. ((2)), there is a region, determined by the Gaussian well, where pair distances occur around $\bar{d}$. In this case, topological contacts are enumerated and eventually they can include some native contacts. One can measure the average of C_t as a function of Q, and $\langle C_t\rangle(Q)$ curve is a signature of the folding mechanisms and depends only on the protein topology. The Mean First Passage Time (MFPT) calculations are performed at a constant temperature and each run is initialized in an open random configuration ($Q_{\text{unf}} = 0.1$). The simulation is performed until it reaches the folded state ($Q_{\text{fold}} = 0.8$). The first passage times are recorded and the MFPT is an average of more than 500 independent runs for each temperature.

RESULTS

This study addresses the effects of frustration in the folding process with the addition of the energy perturbation term [Eq. (2)] to the original unperturbed Hamiltonian [Eq. (1)]. The energetic frustration strength parameter (ε_f) is varied for a set of 19 proteins with different fold motifs and sizes. The results for the perturbed model show that a subset of these characterized proteins has an optimal strength of ε_f that enhances folding and stability. The criterion for determining optimal ε_f is presented in the following section.

Optimal frustration criterion for protein folding

Figure 1(A) shows the folding temperature (T_f) as a function of the energy perturbation strength (ε_f) for a subset of simulated proteins. In Figure 1(A), T_f is normalized by T_f^0, the protein-folding temperature in the absence of frustration [$\varepsilon_f = 0$ in Eq. (2)]. T_f is defined as the temperature where the peak in the specific heat as a function of the temperature is located for each ε_f simulated (Supporting Information Fig. S2). The amount of frustration added to the system contributes to the folding process differently for each protein. In some cases, the addition of a small amount of frustration has no effect on thermodynamic stability (protein PtL) and in other cases T_f decreases for any value of ε_f (protein HP36 and α_3D). However, in some cases, the addition of slight frustration monotonically increases thermodynamic stability until it reaches a maximum defined here as $\varepsilon_f^{\text{opt}}$, which is the optimal frustration strength that maximizes thermodynamic stability of the system at T_f. High values of ε_f always decrease thermodynamic stability and slow kinetics [Fig. 1(B)]. Figure 1(B) shows the folding time (τ_f) versus the frustration parameter ε_f simulated at T_f and normalized by the folding time for the system without frustration (τ_f^0) calculated on T_f^0. Comparing Figure 1(A,B), we observe that an increase in ε_f changes the protein's thermodynamic and kinetics differently. There are cases in which there is no increase in T_f upon the addition of frustration, the kinetics is also slowed (proteins HP36 and α_3 D), and folding time increases monotonically. For the other proteins in Figure 1(B), τ_f decreases, indicating that kinetics is accelerated. The minimum in Figure 1(B) corresponds to ε_f with the most

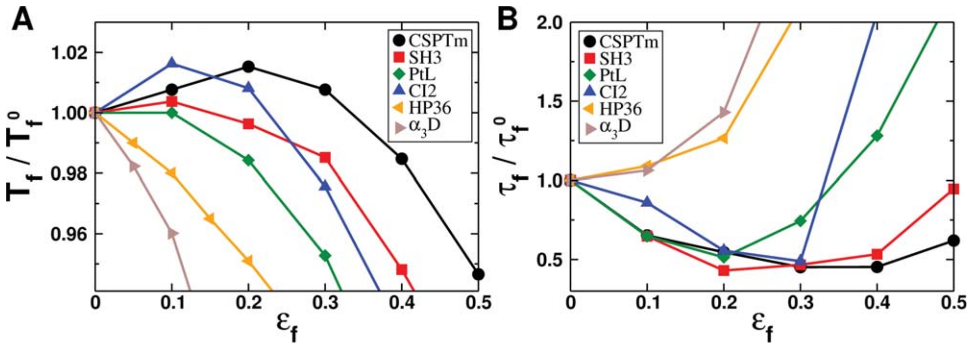

Figure 1

(**A**) Folding temperature (T_f) and (**B**) folding time (τ_f) as a function of the energetic frustration parameter (ε_f) for different proteins. Temperatures in (A) and folding time in (B) are normalized by the respective quantity found in the system without including energy perturbation ($\varepsilon_f = 0$), respectively T_f^0 and τ_f^0. Adding a relatively small amount of frustration ($\varepsilon_f > 0$) increases the thermodynamic stability in (A) and enhances kinetics in (B) only for a subset of the simulated proteins. For this subset, an optimal value of ε_f is identified for thermodynamics and kinetics.

efficient kinetics at T_f, which does not necessarily correspond to the maximum of T_f obtained in Figure 1(A).

The addition of frustration to the model separates the studied proteins into two distinct groups: proteins in which frustration "helps" and those in which frustration hinders the folding event. We defined a criterion that determines when a small amount of frustration is favorable or not for folding. The criterion is based first on thermodynamics: the maximum in T_f/T_f^0 corresponds to the optimum frustration strength ($\varepsilon_f^{\rm opt}$). In the case that there is a maximum in T_f/T_f^0, ε_f presents the fastest kinetics within the plateau are defined as $\varepsilon_f^{\rm opt}$. This simple criterion allows the straightforward determination of the amount of $\varepsilon_f^{\rm opt}$ which maximizes protein thermodynamic stability and kinetic. The values of $\varepsilon_f^{\rm opt}$ for all studied proteins are listed in Table I, along with other quantities relevant to this study.

Strictly speaking, thermodynamic stability is associated with the ratio T_f/T_g, where T_g is the glass temperature, and we did not determine T_g as a function of frustration. The underlying assumption is that T_g is weakly dependent on a small addition of frustration. Indeed, it was shown in C_α models that T_g increases monotonically with energetic frustration, particularly in the low-frustration regime.[35] It is expected that T_g has a similar dependence for all proteins studied here, which causes a monotonic shift in the thermodynamic stability curve [Fig. 1(A)]. This assumption is expected to have a minimal effect on the conclusions regarding the different optimal regimes of frustration.

Thermodynamics and kinetics of optimum frustration regime

Thermodynamics is explored in more detail in Figure 2, which shows the cases where frustration assists and hinders protein folding. The two cases are exemplified by the proteins CI2 and α_3D. Figure 2 shows the two-dimensional free-energy profile $F(Q,C_t)$ as a function of reaction coordinates, native contacts Q and topological contacts C_t, which were defined previously. Each free energy profile is shown at its corresponding T_f and normalized by k_BT_f. Figure 2 shows the effective free-energy barrier ΔF_{2D} between the two valleys, the unfolded (left valley) and folded states (right valley). We use the subscript 2D to distinguish between the free-energy barrier ΔF in the one-dimensional representation of free energy as a function of Q (Table I). Figure 2(A,B) shows

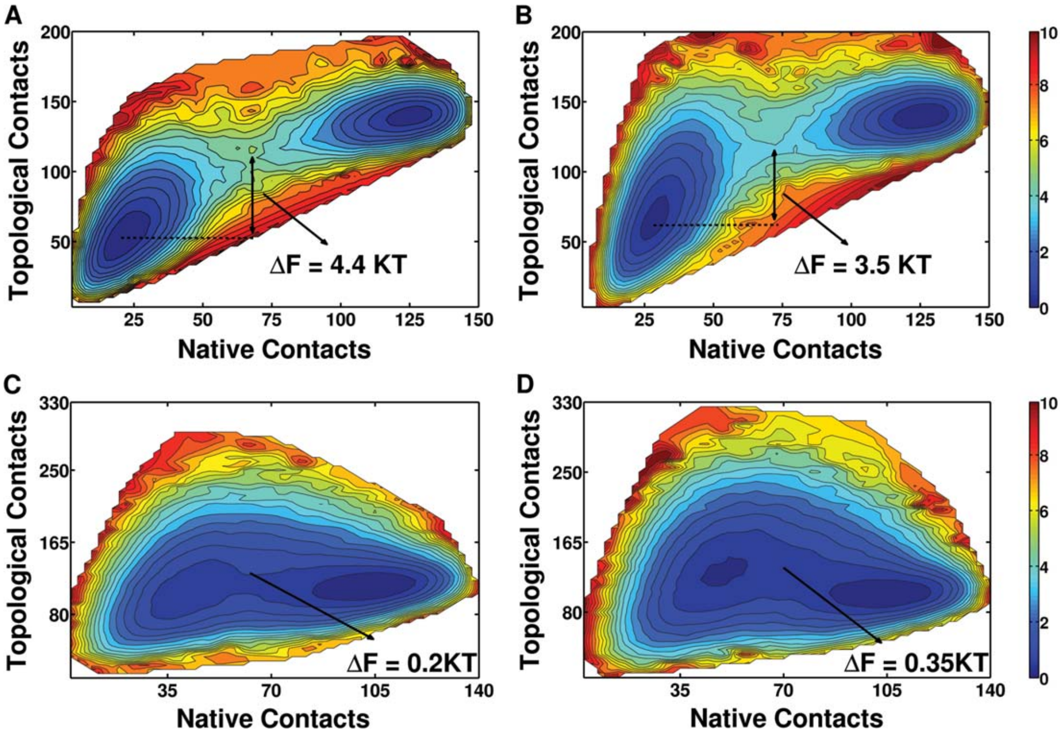

Figure 2

Two-dimensional free-energy profile ($F(Q,C_t)$) as a function of native contacts Q and topological contacts C_t at folding temperature (T_f). $F(Q,C_t)$ is normalized by the corresponding k_BT_f, that is, $F(Q,C_t)$ is in units of k_BT_f. (**A** and **B**) The free energy for CI2 and (**C** and **D**) for α_3D. In (A and C), proteins are unfrustrated ($\varepsilon_f = 0$) and in (B) CI2 has energetic frustration strength ($\varepsilon_f = \varepsilon_f^{\rm opt}$) and the effective free-energy barrier (ΔF) decreases upon the addition of non-native interaction. CI2 is the case that energetic frustration accelerates folding; in (B) folding is faster than in (A). α_3D is the case that energetic frustration hinders folding, in (**D**) folding is slower than in (**C**).

$F(Q,C_t)$ for CI2 and Figure 2(C,D) shows $F(Q,C_t)$ forα_3D. Figure 2(A,C) shows $F(Q,C_t)$ for the proteins without energetic frustration (ε_f = 0.0) and Figure 2(B,D) shows $F(Q,C_t)$ with frustration, ε_f = 0.1 and ε_f = 0.05 for CI2 and α_3D, respectively. In Figure 2(B), CI2 is the case where frustration assists folding to the native state, enhancing kinetics and thermodynamics up to the optimal value of frustration with ΔF_{2D} decreasing. On the other hand, in Figure 2(D), α_3D is the case where folding is slowed down with the addition of any amount of frustration, and ΔF_{2D} increases when compared with the unfrustrated regime in Figure 2(C). The diffusion coefficient D was also calculated as a function of the frustration strength ε_f for different proteins as shown in Supporting Information Figure S3. Essentially, D monotonically decays as energetic frustration increases for all simulated protein which correlates with a slow kinetics as frustration is added.

Figure 3(A,B) shows the average of topological contacts $\langle C_t \rangle$ as a function of Q for different levels of energetic frustration. In the overall shape of the curve, an increase may be observed in $\langle C_t \rangle$ when ε_f increases. For the protein CI2 [Fig. 3(A)], at up to the optimal frustration curve (ε_f^{opt}=0.1), the formation of topological and native contacts is performed in a cooperative manner and the folding reaction speed grows, leading the protein directly to the native state (for the reaction rates, see Fig. 1(B)). For the values of frustration greater than ε_f^{opt}, there is a large occurrence of topological contacts. If topological contacts are excessively formed, it appears as a bump located at the transition state in the $\langle C_t \rangle$ versus Q curve [Fig. 3(A,B)]. These excessive contacts should be disrupted before the protein reaches the folded state. Such behavior contrasts to the mechanism in which there is a monotonic raise of $\langle C_t \rangle(Q)$. The occurrence of excessive contacts can be associated to a slow dynamics, which is supported by higher folding times when compared to those at ε_f^{opt}. The same folding mechanism is observed for the protein α_3D [Fig. 3(B)]. For α_3D, ε_f^{opt}=0.0 and thus any nonzero value of energetic frustration causes the formation of a large quantity of excessive topological contacts, more than in the native state, and therefore, frustration is not favorable to the folding reaction of this protein.

Figure 3(C,D) shows the average of the radius of gyration $\langle R_g \rangle$ as a function of Q with increasing ε_f.

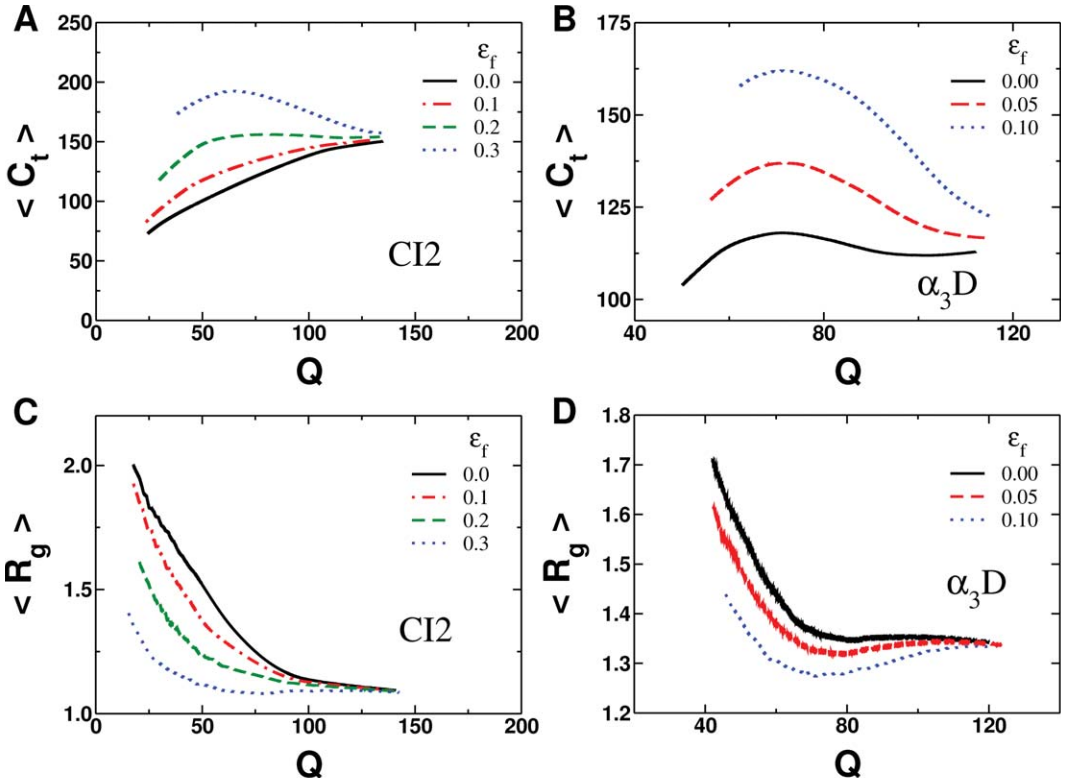

Figure 3

Average of topological contacts ($\langle C_t \rangle$) in (**A** and **B**) and average of the radius of gyration ($\langle R_g \rangle$) in (**C** and **D**) versus native contacts (Q) for different energetic frustration strengths (ε_f) for CI2, (**A** and **C**) and α_3D, (**B** and **D**). The optimum kinetics for CI2 is at $\varepsilon_f=\varepsilon_f^{opt}=0.1$ and α_3D is at $\varepsilon_f^{opt}=0.0$. For CI2, $\langle C_t \rangle$ and $\langle R_g \rangle$ are monotonically dependent on Q up to $\varepsilon_f=\varepsilon_f^{opt}$ where folding is speeded up. For CI2 and α_3D at $\varepsilon_f > \varepsilon_f^{opt}$, $\langle C_t \rangle$ and $\langle R_g \rangle$ are not monotonically dependent on Q and folding is decelerated owing to the excessive formation of C_t.

Comparing with $\langle C_t \rangle$, $\langle R_g \rangle$ also gives the protein compaction degree as folding occurs (increasing Q). R_g is more studied by the scientific community than C_t and can be directly measured by experimentalists. Similar analysis to that in Figure 3(A,B) is performed in Figure 3(C,D), and the same folding mechanism is provided. Figure 3(C) shows, for protein CI2, the $\langle R_g \rangle$ profile without energetic frustration ($\varepsilon_f = 0.0$) and with optimal energetic frustration ($\varepsilon_f^{\mathrm{opt}}$=0.1). Optimal energetic frustration induces compact states denser than in the case without energetic frustration, resulting in accelerated formation of Q. This process leads to a fast folding rate, the same effect shown in Figure 3(A). For protein $\alpha_3 D$ shown in Figure 3(D), energetic frustration ($\varepsilon_f = 0.05$) induces more compact states during folding with $\langle R_g \rangle$ close to the folding transition state ($Q = 70$) smaller than $\langle R_g \rangle$ in the native state ($Q = 70$). In other words, to fold, the protein collapses ($\langle R_g \rangle$ decreases before the transition state) and, after that, the protein is required to increase its R_g, breaking wrong contacts, and then cross the folding barrier, reaching the native basin ($\langle R_g \rangle$ remains almost constant after the transition state). Energetic frustration in the case of $\alpha_3 D$ delays folding owing to the wrong topological contacts formed prematurely, the same mechanism supported by Figure 3(B). The increase of R_g compaction upon addition of non-native interactions [Fig. 3(C,D)] was also reported recently in Ref. 56 in which the authors used a potential with native contacts between pairs of C_α atoms that have been formed and they included electrostatic and non-native interactions in addition to the native-centric potential, one observes the propensities of knotting events in the early stages of folding owing to the addition of non-native interactions.

Accelerated folding rates and decreased free-energy barriers in optimum energetic frustrated protein were previously reported by theory and computational experiments.[26,27,34] These reports correspond to the cases of small globular proteins in which a little frustration always enhances folding up to its optimal degree. Clementi and Plotkin[27] associated these effects with the assumption that the native topological constraints induce a protein collapse at the non-native part of the structure. This mechanism is better understood if we think in terms of energetic frustration and topological contacts. Energetic frustration is a perturbative and attractive interaction which could induce protein collapse. We show that, for some proteins, a small quantity of energetic frustration accelerates native contact formation during the folding reaction, which has the effect of increasing the folding rate up to a certain value of energetic frustration. If frustration increases beyond the optimal level of frustration, nonspecific attractive forces between residues become dominant. An excess of topological contacts is formed and the folding rate decreases. For some proteins, this optimal level of frustration is the unperturbed system.

The addition of a small amount of non-native interaction lowers the free-energy barrier at the transition state. Oliveira *et al.*[34] found that a low level of frustration stabilizes the native ensemble and increases the diffusion of the protein to the folded state. Using ϕ-value analysis, a more homogeneous structural content at the transition state ensemble was also observed. Thus, the decrease in the free-energy barrier upon energetic frustration could be associated with this homogeneous formation of native contacts in the transition state ensemble and the increase of the internal friction shown by the overall decrease on the diffusion (Supporting Information Fig. S3). These effects on the free-energy barrier have been measured in an ultrafast folding protein[57] and experimentally observed in some systems like spectrins.[58,59]

Classification of proteins according to optimal frustration criterion

In general, non-native energetic frustration influences the system in two different opposite ways: accelerating or hindering folding speed. Here, we classify all 19 proteins simulated into these two groups: one in which non-native energetic frustration helps ($\varepsilon_f^{\mathrm{opt}} > 0$), and another in which any energetic frustration hinders folding ($\varepsilon_f^{\mathrm{opt}}$=0). One relevant question is how one could predict in which group a given protein belongs solely on a simple characterization, such as the parameters listed in Table I. We observed significant correlations between the free-energy barrier (ΔF) and the ACO with the optimal energetic frustration strength ($\varepsilon_f^{\mathrm{opt}}$) for our protein data set, which was expected as ΔF and ACO are strongly correlated.[8] However, the distinction between the two groups, when looking at each variable alone, is not so striking, by which we mean that, if we change ΔF or ACO by less than a factor of 2, it may be enough to swap from one group to another.

Figure 4 shows ΔF versus ACO for the proteins with their respective level of $\varepsilon_f^{\mathrm{opt}}$ as summarized in Table I. We clearly distinguish two clusters of proteins: one group has a relatively high free-energy barrier and a high absolute contact order (red delimited region, Fig. 4) with proteins corresponding to the case where little energetic frustration assists folding ($\varepsilon_f^{\mathrm{opt}} > 0$). The second cluster has a relatively low ΔF and low ACO (blue region, Fig. 4) and characterizes proteins where energetic frustration does not favor folding ($\varepsilon_f^{\mathrm{opt}}$=0). The occurrence of these two groups does not correlate with protein size.

With regard to protein motif, Figure 4 also shows the structural classification of proteins (SCOP) database criterion.[60] Figure 4 has proteins belonging to the three different SCOP motifs: α (circles), $\alpha + \beta$ (diamonds), and β (triangles). In Figure 4, the blue delimited group with $\varepsilon_f^{\mathrm{opt}}$=0.0 (which we refer to as naturally optimized protein) has only proteins with α-motif, and the red group with $\varepsilon_f^{\mathrm{opt}} > 0.0$ (computationally optimized group)

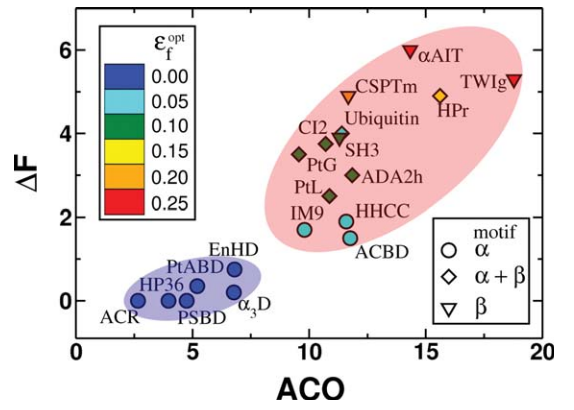

Figure 4

Free-energy barrier (ΔF) as a function of ACO for all proteins studied. Proteins are colored by their respective optimum energetic frustration degree (ε_f^{opt}) and each protein is also represented by its respective SCOP database criterion[60]: α (circles), $\alpha + \beta$ (diamonds), and β (triangles). Two distinct groups of proteins are delimited by filled ellipses. One group shows lower ΔF and ACO with ε_f^{opt}=0.0 (blue shaded area), for this group, folding is hindered upon the addition of ε_f and it has only proteins with α-motif. The other group shows higher ΔF and ACO with $\varepsilon_f^{opt} > 0$ (red shaded area) and has the three protein motifs (α, $\alpha + \beta$, and β).

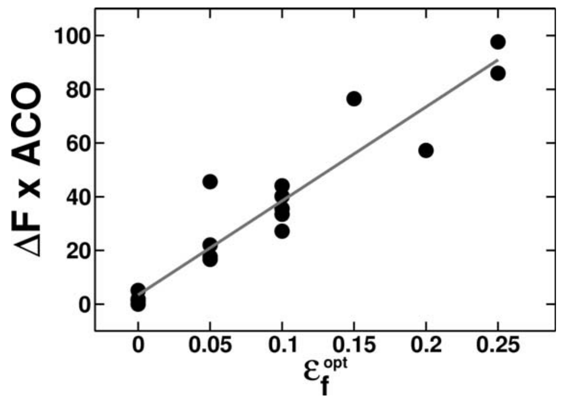

Figure 5

Free-energy barrier times and ACO ($\Delta F \times$ ACO) as a function of the optimal energetic frustration strength ε_f^{opt}. Each point represents one simulated protein. The linear fit correlation to the data is 0.95.

has the three protein motifs (α, $\alpha + \beta$, and β). We could speculate, by inspecting these results, that in general, β proteins are those that could have their kinetics optimized by select mutations that create little energetic frustration. Evolution has selected α-proteins to be naturally optimized. $\alpha + \beta$-Proteins could be the middle step in this evolutionary step and would require even less energetic frustration than β-proteins to have faster kinetics.

It seems that, as in the case of the protein size, the protein fold motif does not separate proteins in the two regions delimited by the nonfrustrated (blue) and frustrated (red) regimes as shown in Figure 4. We aim to find one single parameter that is able to predict which protein is naturally energetically optimized and which one could be experimentally enhanced, as well as to establish a robust threshold to separate the two regimes identified in this study without having to perform massive high-performance computation.

The simplest nonlinear combination of these parameters is the product between the energetic parameter ΔF and the structural parameter ACO. Figure 5 shows the product $\Delta F \times$ ACO versus ε_f^{opt} for the protein data (Table I). The linear fit for this plot has a very strong correlation (0.95). Moreover, there is a large gap in $\Delta F \times$ ACO as a function of ε_f^{opt}; the product increases by at least a factor of 3 (from 5.1 to 16.7) from one group to another. Thus, the high correlation and gap shows that the product $\Delta F \times$ ACO is the candidate parameter that will serve as the threshold reference for predicting the intrinsically energetic frustrated proteins.

A single variable (size or motif, ΔF, ACO), is not sufficient to classify or even predict the intrinsic energetic protein frustration. However, the product $\Delta F \times$ ACO can successfully classify the two groups. These results are corroborated by principal component analysis (PCA) and partial least squares (PLS) calculations.

PCA and PLS

The correlation between $\Delta F \times$ ACO and ε_f^{opt} was inferred by inspection. A more systematic multivariate analysis was carried out[61–64] using the entire data of Table I. Initially, PCA was performed. Figure 6(A) shows the PCA scores for the protein data set using variables from Columns 3 to 8 (ε_f^{opt} information was not included). It must be stressed that this classification does not depend on the way frustration is added to potential energy, it uses only the information based on the structure-based model. Other models for non-native interaction potential also report optimal folding mechanisms.[28–30,37,38,56] With two PCAs, 93.5% of the whole information was accounted for. These results show that the two aforementioned groups are completely distinguishable. According to the loadings diagram [Fig. 6(B)], $\bar{d}$ is the only variable that presents a negative coefficient in the first PC.

For PLS results, Figure 7 shows the regression of all variables against variable ε_f^{opt}, considered as the dependent variable. Two principal components were necessary to perform the PLS regression, and the values listed in Table II show that the standard error for validation (SEV) and standard error for calibration (SEC) are minima for two PCs. The internal (Q^2) and external (R^2) regression coefficients are 0.891 and 0.927, respectively, indicating a high level of correlation. Complementary PCA and PLS analyses were performed without $\bar{d}$, a parameter which is dependent on the defined contact map,

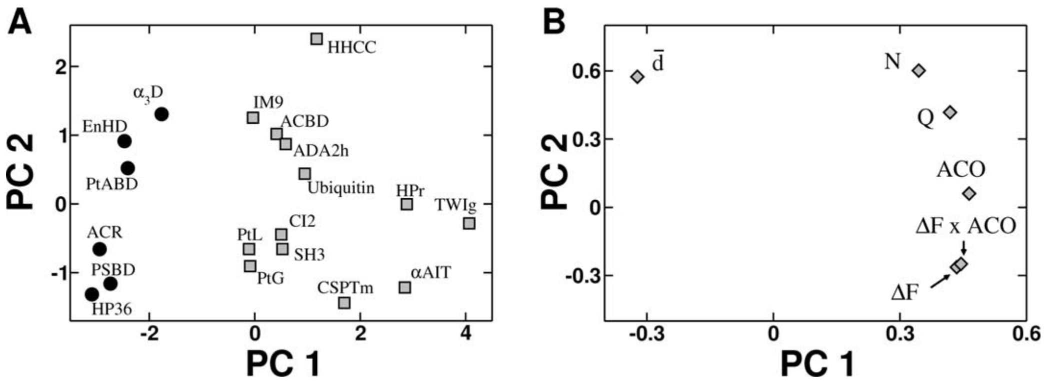

Figure 6

(**A**) PCA scores for the protein data set, indicating the segregation between the same two groups shown in Figure 4. Blue circles show the proteins that $\varepsilon_f^{opt}=0.0$ and red squares show the proteins that $\varepsilon_f^{opt} > 0.0$. (**B**) Loadings diagram of the used variables from Columns 3 to 8 in Table I.

and the results were similar to those shown in Figures 6 and 7 (Supporting Information Figs. S4 and S5).

DISCUSSION AND CONCLUSIONS

The effect of a small amount of non-native energetic frustration on the folding reaction in 19 proteins different in size and motif was demonstrated. Proteins can be well separated into two groups: in one group, a small amount of energetic frustration helps to accelerate the folding process. In the other group, folding is always hindered by any quantity of energetic frustration added to the system. Folding transition temperature (T_f) and kinetic rates (τ_f) are enhanced for some proteins and decreased for others upon the addition of energetic frustration There is a very strong correlation between the free-energy barrier times absolute contact order ($\Delta F \times$ ACO) as a function of the optimal energetic frustration value (ε_f^{opt}).

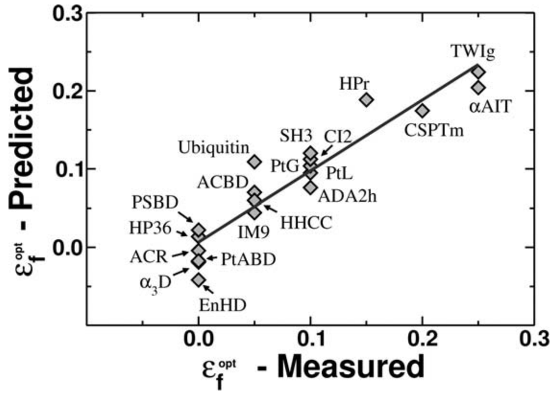

Figure 7

Results of PLS using variables from Table I, from Columns 3 to 8, against ε_f^{opt}. Measured values are those in Column 9 of Table I and predicted values were calculated using the PLS model.

The structure-based model perturbed by energetic frustration was shown to be a robust computational model to simulate relatively small globular proteins and to indicate in which cases proteins should be a little more energetically frustrated. To verify these statements established by simulations, PCA and PLS analyses were performed using the properties extracted from the protein data set (Table I). PCA and PLS analyses segregate the proteins into the same two groups from the simulation results. This is the evidence that our minimalistic model, with an energetic frustration homogenously added to the system, is able to predict which proteins could be optimized. Also, PCA and PLS support the conclusion that the simulation results reported are not dependent on the non-native interaction form of the potential energy function [Eq. (2)], the extra term of the native-centric structure-based model.

These predictions should be able to guide experimentalists to select the proteins whose thermodynamic stability and folding rates could be enhanced. This could be achieved by mutating residues where specific sites could be energetically perturbed. As local frustration is often associated with function and can be, in principle, computationally calculated,[31] the model proposed here

Table II

PLS Results Using All Variables from Table I Against ε_f^{opt}

	Variance	Percent	Cumulative	SEV[a]	Q^2	SEC[b]	R^2
Factor 1	77.340	71.611	71.611	0.035	0.803	0.032	0.854
Factor 2	**23.530**	**21.787**	**93.398**	**0.0264**	**0.891**	**0.023**	**0.927**
Factor 3	2.320	2.149	95.546	0.029	0.865	0.0235	0.931
Factor 4	3.589	3.323	98.870	0.030	0.859	0.0240	0.935

[a] Standard error in validation.
[b] Standard error in calibration.

brings direct insights to experimental studies on protein-folding mechanisms.

Results for C_α structure-based model show that T_f/T_g behavior scales with T_f and in the optimum frustration regime might have slightly higher T_f/T_g when compared with the unfrustrated system.[35] This increase should be negligible, and also in practical terms, C_α structure-based models are unrealistically unfrustrated, such that, even in the regime in which some energetic frustration accelerate folding, the observed energetic frustration in a real system is probably higher than the optimum regime estimated here. Nevertheless, one can clearly distinguish between the two contrasting behaviors: one in which frustration does not significantly affect stability and improve folding rates and another in which frustration is detrimental to folding.

High ΔF and high ACO might be associated with high topological frustration, which provides conditions for $\varepsilon_f^{\text{opt}} > 0$, under which some energetic frustration can partially assist folding (mostly $\alpha + \beta$ and β-proteins). On the other hand, low ΔF and low ACO are related to proteins evolved to a naturally optimized state ($\varepsilon_f^{\text{opt}}=0$) with their energetic and entropic parts, from ΔF, balanced with their topological restraints in such a way that no more energetic frustration is required by evolution (mostly, α-proteins).

In this study, we have carried out analysis using only computational data. Although ΔF can be extracted experimentally, such as inferred from chevron plots obtained by kinetic spectroscopy probes, it will be interesting to test the ideas derived here in real systems.

ACKNOWLEDGMENTS

The authors thank Jin Wang, Paul Whitford, and Leandro Oliveira for helpful discussions. The authors also thank GridUNESP for the computational resources.

REFERENCES

1. Levinthal C. Mossbauer spectroscopy in biological systems. In: Debrunner P, Tsibris J, Munch E, editors. Proceedings of a meeting held at Allerton house, Monticello, Illinois. Urbana: University of Illinois Press; 1969. pp 22–24.
2. Leopold PE, Montal M, Onuchic JN. Protein folding funnels—a kinetic approach to the sequence structure relationship. Proc Natl Acad Sci USA 1992;18:8721–8725.
3. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein-folding—a synthesis. Proteins 1995;21:167–195.
4. Wolynes PG, Onuchic JN, Thirumalai D. Navigating the folding routes. Science 1995;267:1619–1620.
5. Dill KA, Chan HS. From Levinthal to pathways to funnels. New J Phys 1997;4:10–19.
6. Chan HS, Zhang Z, Wallin S, Liu Z. Cooperativity, local-nonlocal coupling, and non-native interactions: principles of protein folding from coarse-grained models. Annu Rev Phys Chem 2011;62:301–326.
7. Nymeyer H, Garcia AE, Onuchic JN. Folding funnels and frustration in off-lattice minimalist protein landscapes. Proc Natl Acad Sci USA 1998;95:5921–5928.
8. Koga N, Takada S. Roles of native topology and chain-length scaling in protein folding: a simulation study with a gö-like model. J Mol Biol 2001;313:171–180.
9. Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu K, Onuchic JN. An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. Proteins: Struct Funct Bioinform 2009;75:430–441.
10. Fersht AR. Characterizing transition states in protein folding: an essential step in the puzzle. Curr Opin Struct Biol 1995;5:79–84.
11. Garcia-Mira MM, Sadqi M, Fischer N, Sanchez-Ruiz JM, Muoz V. Experimental identification of downhill protein folding. Science 2002;298:2191–2195.
12. Nettels D, Gopich IV, Hoffmann A, Schuler B. Ultrafast dynamics of protein collapse from single-molecule photon statistics. Proc Natl Acad Sci USA 2007;104:2655–2660.
13. Chung HS, Louis JM, Eaton WA. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. Proc Natl Acad Sci USA 2009;106:11837–11844.
14. Chavez LL, Onuchic JN, Clementi C. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. J Am Chem Soc 2004;126:8426–8432.
15. Snow CD, Sorin EJ, Rhee YM, Pande VS. How well can simulation predict protein folding kinetics and thermodynamics? Annu Rev Biophys Biomol Struct 2005;34:43–69.
16. Bryngelson JD, Wolynes PG. Spin-glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84:7524–7528.
17. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science 1991;254:1598–1603.
18. Socci ND, Onuchic JN, Wolynes PG. Diffusive dynamics of the reaction coordinate for protien folding funnels. J Chem Phys 1996;104:5860–5868.
19. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. Toward an outline of the topography of a realistic protein-folding funnel. Proc Natl Acad Sci USA 1995;92:3626–3630.
20. Anfinsen CB. Principles that govern the folding of proteins chains. Science 1973;181:223–230.
21. Onuchic JN, Nymeyer H, Garcia AE, Chahine J, Socci ND. The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. Adv Protein Chem.2000;53:87–152.
22. Shakhnovich EI, Gutin AM. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. Biophys Chem 1989;34:187–199.
23. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy-model (with applications to protein folding). J Phys Chem 1989;93:6902–6915.
24. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992;89:4918–4922.
25. Sutto L, Ltzer J, Hegler JA, Ferreiro DU, Wolynes PG. Consequences of localized frustration for the folding mechanism of the IM7 protein. Proc Natl Acad Sci USA 2007;104:19825–19830.
26. Plotkin SS. Speeding protein folding beyond the Gō model: how a little frustration sometimes helps. Proteins: Struct Funct Genet 2001;45:337–345.
27. Clementi C, Plotkin SS. The effects of non-native interactions on protein folding rates: theory and simulation. Prot Sci 2004;13:1750–1766.
28. Zarrine-Afsar A, Wallin S, Neculai AM, Neudecker P, Howell PL, Davidson AR, Chan HS. Theoretical and experimental demonstration of the importance of specific nonnative interactions in protein folding. Proc Natl Acad Sci USA 2008;105:9999–10004.
29. Zarrine-Afsar A, Zhang Z, Schweiker KL, Makhatadze GI, Davidson AR, Chan HS. Kinetic consequences of native state optimization of

surface-exposed electrostatic interactions in the fyn SH3 domain. Proteins: Struct Funct Bioinform 2012;80:858870.

30. Shental-Bechor D, Smith MTJ, MacKenzie D, Broom A, Marcovitz A, Ghashut F, Gö C, Bralha F, Meiering EM, Levy Y. Nonnative interactions regulate folding and switching of myristoylated protein. Proc Natl Acad Sci USA 2012;109:17839–17844.
31. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG. Localizing frustration in native proteins and protein assemblies. Proc Natl Acad Sci USA 2007;104:19819–19824.
32. Gosavi S, Chavez LL, Jennings PA, Onuchic JN. Topological frustration and the folding of interleukin-1 beta. J Mol Biol 2006;357:986–996.
33. Plotkin SS, Onuchic JN. Understanding protein folding with energy landscape theory. Part i: basic concepts. Q Rev Biophys 2002;35:111–167.
34. Oliveira RJ, Whitford PC, Chahine J, Wang J, Onuchic JN, Leite VBP. The origin of nonmonotonic complex behavior and the effects of nonnative interactions on the diffusive properties of protein folding. Biophys J 2010;99:600–608.
35. Wang J, Oliveira RJ, Chu X, Whitford PC, Chahine J, Han W, Wang E, Leite VBP. The topography of funneled landscapes determines the thermodynamics and kinetics of protein folding. Proc Natl Acad Sci USA 2012;109:15763–15768.
36. Shea J, Onuchic JN, Brooks CL. Energetic frustration and the nature of the transition state in protein folding. J Chem Phys 2000;113:7663–7671.
37. Zhang Z, Chan HS. Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins. Proc Natl Acad Sci USA 2010;107:2920–2925.
38. Azia A, Levy Y. Nonnative electrostatic interactions can modulate protein folding: molecular dynamics with a grain of salt. J Mol Biol 2009;393:527–542.
39. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
40. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298:937–953.
41. Ueda Y, Taketomi H, Gō N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effects of specific amino acid sequence represented by specific inter-unit interactions. Int J Peptide Res 1975;7:445–459.
42. Sobolev V, Wade R, Vried G, Edelman M. Molecular docking using surface complementarity. Proteins: Struct Funct Genet 1996;25:120–129.
43. Matouschek A, Kellis JT, Serrano L, Fersht AR. Mapping the transition state and pathway of protein folding by protein engineering. Nature 1989;340:122–126.
44. Yang S, Onuchic JN, Levine H. Effective stochastic dynamics on a protein folding energy landscape. J Chem Phys 2006;125:054910–054918.
45. Best RB, Hummer G. Coordinate-dependent diffusion in protein folding. Proc Natl Acad Sci USA 2010;107:1088–1093.
46. Chahine J, Oliveira RJ, Leite VBP, Wang J. Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. Proc Natl Acad Sci USA 2007;104:14646–14651.
47. Oliveira RJ, Whitford PC, Chahine J, Leite VBP, Wang J. Coordinate and time-dependent diffusion dynamics in protein folding. Methods 2010;52:91–98.
48. Calculate the contact order of proteins: Baker laboratory, department of biochemistry, university of Washington. URL: http://depts.washington.edu/bakerpg/contact_order/.
49. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
50. Plaxco KW, Simons KT, Ruczinski I, Baker D. Topology, stability, sequence, and length: defining the determinants of Two-State protein folding kinetics. Biochemistry 2000;39:11177–11183.
51. Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN. SMOG@ctbp: simplified deployment of structure-based models in GROMACS. Nucleic Acids Research 2010.
52. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: fast, flexible, and free. J Comp Chem 2005;26:1701–1718.
53. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684.
54. Ferrenberg AM, Swendsen RH. New Monte Carlo technique for studying phase transitions. Phys Rev Lett 1988;61:2635–2638.
55. Cho SS, Levy Y, Wolynes PG. P versus Q: structural reaction coordinates capture protein folding on smooth landscapes. Proc Natl Acad Sci USA 2006;103:586–591.
56. Škrbić T, Micheletti C, Faccioli P. The role of non-native interactions in the folding of knotted proteins. PLoS Comput Biol 2012;8:e1002504.
57. Cellmer T, Henry ER, Hofrichter J, Eaton WA. Measuring internal friction of an ultrafast-folding protein. Proc Natl Acad Sci USA 2008;105:18320–18325.
58. Borgia A, Wensley BG, Soranno A, Nettels D, Borgia MB, Hoffmann A, Pfeil SH, Lipman EA, Clarke J, Schuler B. Localizing internal friction along the reaction coordinate of protein folding by combining ensemble and single-molecule fluorescence spectroscopy. Nat Commun 2012;3:1195.
59. Wensley BG, Kwa LG, Shammas SL, Rogers JM, Browning S, Yang Z, Clarke J. Separating the effects of internal friction and transition state energy to explain the slow, frustrated folding of spectrin domains. Proc Natl Acad Sci USA 2012;109:17795–17799.
60. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
61. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics Intell Lab Syst 1987;2:37–52.
62. Beebe KR, Pell RJ, Seasholtz MB. Chemometrics: a practical guide, 1st ed. New York: Wiley-Interscience; 1998.
63. Lavine B, Workman, JJ, Jr. Chemometrics. Anal Chem 2004;76:3365–3371.
64. Miller PJ, Miller JC. Statistics and chemometrics for analytical chemistry, 6th ed. Harlow, England: Prentice Hall; 2010.