13

Estimation of Folding Probabilities and Φ Values From Molecular Dynamics Simulations of Reversible Peptide Folding

Francesco Rao, Giovanni Settanni, and Amedeo Caflisch

Summary

Molecular dynamics simulations with an implicit model of the solvent have allowed to investigate the reversible folding of structured peptides.

For a 20-residue antiparallel β-sheet peptide, the simulation results have revealed multiple folding pathways. Moreover, the conformational heterogeneity of the denatured state has been shown to originate from high enthalpy, high entropy basins with fluctuating non-native secondary structure, as well as low enthalpy, low entropy traps. An efficient and simple approach to estimate folding probabilities from molecular dynamics simulations has allowed to isolate conformations in the transition state ensemble and to evaluate Φ values, i.e., the effects of mutations on the folding kinetics and thermodynamic stability. These molecular dynamics studies have provided evidence that, if interpreted by neglecting the non-native interactions, Φ values overestimate the amount of native-like structure in the transition state.

Key Words: Protein folding; energy landscape; transition state ensemble; denatured state ensemble; implicit solvent molecular dynamics.

1. Introduction

Energy landscape theory provides a framework for the description of the kinetics and thermodynamics of condensed phases. In the past years, it has been extensively applied to the analysis of protein folding ***(1–5)***. Although proteins are essential macromolecules for life and are responsible for most cellular functions, the process by which proteins reach their functional structure are not fully understood ***(6)***. Within the energy landscape framework, protein folding is envisioned to proceed along a moderately rough funnel-shaped effective energy surface ***(2,7)***. The overall shape of the landscape arises from a strong energetic driving force to

From: *Methods in Molecular Biology, vol. 350: Protein Folding Protocols*
Edited by: Y. Bai and R. Nussinov © Humana Press Inc., Totowa, NJ

the native global minimum. This energetic bias is necessary to overcome the conformational search problem associated with finding the native state of the protein within a biologically reasonable time frame* ***(2,8)***. The roughness of the surface is determined by local energy minima arising from the many competing interactions that are possible between the residues. Energetic traps are sequence-related and arise when non-native but stabilizing contacts form as the chain folds. The number and depth of such energetic traps influence both the thermodynamic and kinetic aspects of folding.

Experimental data ***(9)*** indicate that folding for many small proteins is a first-order transition in which the polypeptide chain passes from a free energy basin associated with low order and mainly stabilized by entropy to a free energy basin characterized by a highly ordered dominant conformation of the chain and mainly stabilized by favorable intraprotein interactions. The conformations populating the barrier dividing the two main free energy basins constitute the transition state ensemble (TSE). Understanding the characteristics of the TSE will allow the identification of the events that determine the folding rate and more in general the folding process itself. For this reason many studies have tried to characterize the TSE of proteins. Experimental data on the TSE of proteins have been mainly obtained by a widely diffused technique known as Φ-value analysis ***(10)***. This technique consists of measuring the change of the height of the free energy barrier relative to the change in stability upon a single-point mutation. The denatured state is taken as reference. In this way it is possible to estimate the amount of native structure in the TSE around the mutated residue. This mainly energetic information however does not provide the atomic resolution that one would like to reach and the interpretation of the experiments is not always straightforward, as will be explained next.

Molecular dynamics (MD) is a very useful simulation approach to study the flexibility of proteins at atomic level of detail ***(11,12)***. Since the first MD simulation of a protein in vacuo published in 1977 ***(13)***, much progress has been made to increase the accuracy of the models and reliability of the simulations. Moreover, computer performances have evolved dramatically. However, even for a small protein it is not yet feasible to simulate reversible folding with a high-resolution approach, e.g., MD simulations with an all-atom transferable model. In this chapter we will show that despite their limitations, computer simulations are an important tool for the investigation of the energy landscapes governing protein folding.

The characterization of the TSE of protein folding has attracted the attention of many theoretical and computational studies ***(14–19)***. By definition,

*In contrast with the astronomical amount of time needed by a random search in the configuration space of the protein (Levinthal's paradox).

TSE conformations have a 50% probability of reaching the folded state before unfolding (p_{fold}). Because p_{fold} is computationally very expensive, often the TSE of proteins has been identified on the basis of the projections of the phase space of the protein onto one or two order parameters, i.e., by selecting structures belonging to poorly populated regions of projected free energy landscapes in between the highly populated folded and unfolded state. In what follows we will show the possible problems related to this approach and will present a technique that has been developed to estimate the folding probability of the structures sampled along equilibrium folding-unfolding MD simulations. Such technique has been used to characterize the TSE of folding of the structured peptide Beta3s, a designed 20-residue, three-stranded antiparallel β-sheet ***(20,21)***. We will also discuss how Φ values have been measured *in silico* for this peptide and how their structural interpretation matches the TSE obtained using p_{fold}.

2. Methods

2.1. Molecular Dynamics Simulations

All simulations and part of the analysis of the trajectories were performed with the program CHARMM ***(11)***. Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field ***[11]***). A mean field approximation based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute ***(22)***.

2.2. Clusterization

The 500,000 conformations saved along the 10-μs simulation time of Beta3s ***(23)*** were clustered by the leader algorithm ***(24)***. Briefly, the first structure defines the first cluster and each subsequent structure is compared with the set of clusters found so far until the first similar structure is found. If the structural deviation (*see* below) from the first conformation of all of the known clusters exceeds a given threshold, a new cluster is defined. The leader algorithm is very fast even when analyzing large sets of structures like in the present work. The results presented here were obtained with a structural comparison based on the Distance Root Mean Square (DRMS) deviation considering all distances involving C_α and/or C_β atoms and a cutoff of 1.2 Å. This yielded 78,183 clusters for Beta3s. The DRMS and root mean square deviation of atomic coordinates (upon optimal superposition) have been shown to be highly correlated ***(16)***. The DRMS cutoff of 1.2 Å was chosen on the basis of the distribution of the pairwise DRMS values in a subsample of the wild-type trajectories. The distribution shows two peaks that originate from intra- and intercluster distances. The cutoff is located at the minimum between the first and second DRMS peak.

The main findings of this chapter are valid also for clusterization based on secondary structure similarity ***(19,25)***.

2.3. Definition of TSE

Each cluster i contains $n_f(i)$ snapshots committed to fold out of its total number of snapshots $N(i)$. A cluster j belonging to TSE by definition has an asymptotic $P_f = 0.5$, i.e., if we could extend our simulations so that $N(j) \to \infty$ then $n_f(j)/N(j) \to 0.5$. This means that, if the commitment of each snapshot is considered as an independent binary variable (i.e., 1 or 0 , for a snapshot committed to fold or not, respectively), then the number $n_f(j)$ of snapshots with commitment 1 in a cluster belonging to TSE will follow a binomial distribution with probability $p = 0.5$:

$$P_{N(j)}(n_f(j)) = \binom{N(j)}{n_f(j)} p^{n_f(j)}(1-p)^{(N(j)-n_f(j))} = \binom{N(j)}{n_f(j)} 0.5^{N(j)} \quad (1)$$

Thus, it can be tested if cluster X belongs to TSE by checking that $n_f(X)$ is compatible with a binomial distribution with $p = 0.5$, i.e., $n_f(X)$ has to belong to a likelihood range of values centered around $N(X)/2$. This is done by verifying that the probability to have a hypothetical number n of fold-committed snapshots outside of the range from $n > m(X)$ to $N(X)–n_f(X)$ (that is twice the probability to have $n > m(X) = max(n_f(X),N(X) - n_f(X))$ is larger than a given likelihood confidence threshold λ (e.g., $\lambda = 0.2$ to allow for clusters with three snapshots to belong to the TSE if $n_f = 1$ or $n_f = 2$ because, in both cases, $2 \cdot 1/8 > 0.2$). In mathematical terms:

$$X \in \text{TSE} \Leftrightarrow \sum_{i=m(X)+1}^{N(X)} 2 \cdot P_{N(X)}(i) > \lambda \quad (2)$$

In practice, the latter condition allows TSE clusters with few snapshots to have a larger spread of P_f^C (*see* below for P_f^C definition) around 0.5 than large TSE clusters, because in the approximation of the actual P_f (*see* **Subheading 4.2.**) by P_f^C the error is larger for smaller cluster size.

3. Projection of the Free Energy Landscape on Order Parameters

A common way to investigate and display the free energy landscape is to study it as a function of one or more *order parameters*, i.e., suitably chosen macroscopic quantities that distinguish the different states of the protein. For example, it is common in the study of protein folding to use the fraction of native contacts Q ***(21,26)***. Q is a good *order parameter* in the sense that it distinguishes the unfolded from the folded state: unfolded conformations typically

have small Q, while by definition Q is close to 1 in the native state. The free energy of a protein as a function of Q can be written as:

$$F(Q) = U(Q) - TS(Q) \tag{3}$$

where $F(Q)$, $U(Q)$ and $S(Q)$ are the average free energy, potential energy, and configurational entropy, respectively for the configurations with Q native contacts.

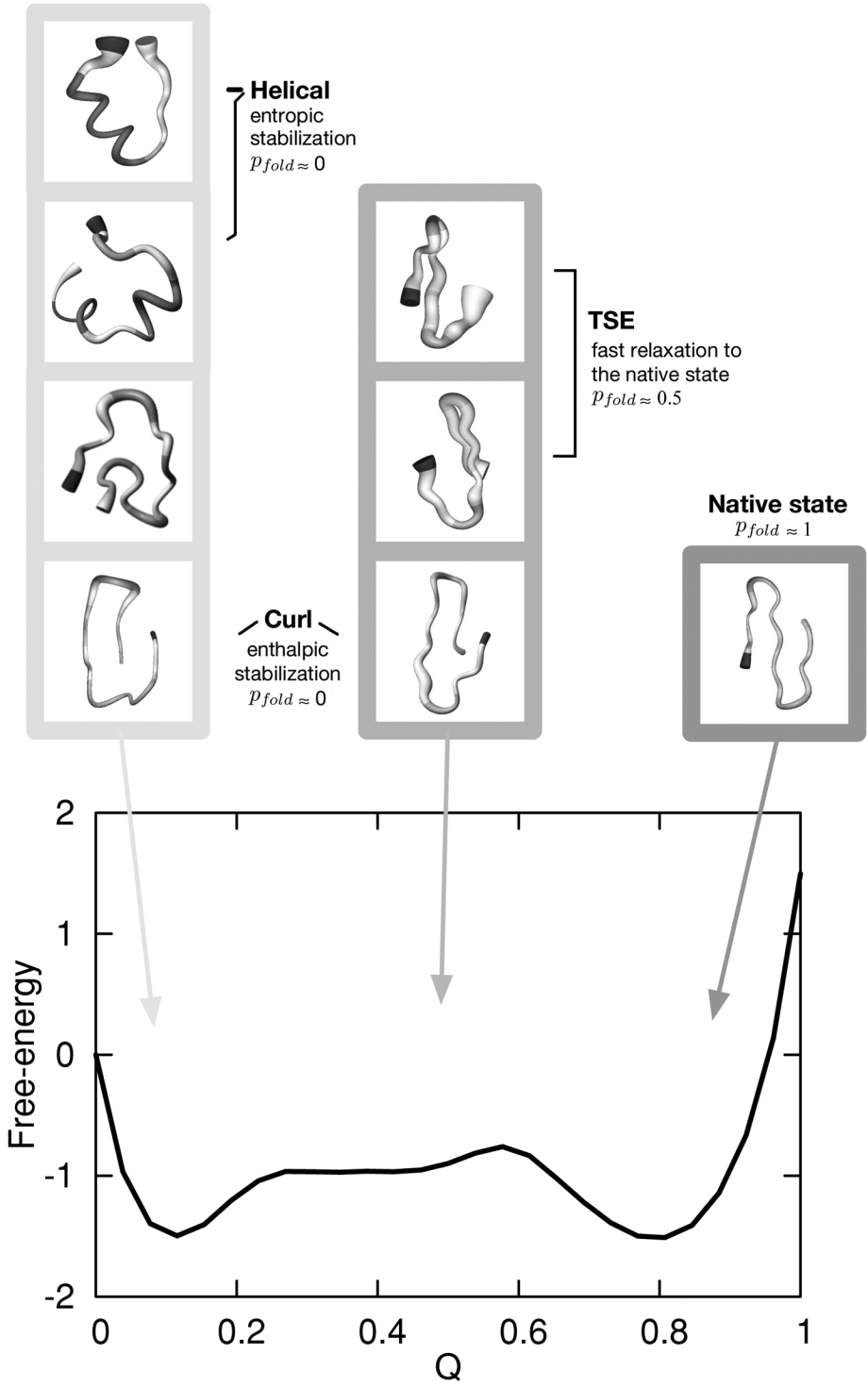
Free energy projections on order parameters have been used to analyze many aspects of protein folding. Stable *states* are associated with local free energy minima of the projected landscape. The depth of the minima is considered proportional to the stability of the states associated to them and the barriers between different minima indicate activation energies between states. In many cases, this approach reveals a surprisingly *simple* two-state picture for protein folding (**Fig. 1**, bottom).

Order parameters are also used as reaction coordinates to monitor the dynamics of the protein ***(14)***. However, using free energy projections for the study of the kinetics of protein folding requires knowledge of a good reaction coordinate, which is not easily accessible and/or identifiable ***(27)***. Given the complexity of protein folding and the large number of degrees of freedom involved, few simple reaction coordinates would be desirable for its description, even though they might miss essential aspects of the process ***(19,26,28)***. Good reaction coordinates for studying the kinetics of protein folding should satisfy two assumptions:

1. The order parameter(s) should allow to distinguish the various states of the system.
2. Within a minimum of the projected free energy landscape, conformations should interconvert rapidly.

A first consequence of assumption 1 is that every value of the order parameter (or the combination of different order parameters) identifies only one state of the system. Assumption 2, stated in a different way, says that all the conformations in a state are kinetically homogeneous. In many cases at least one of these assumptions is not true. For example, if Q is used as order parameter, the conformations with half of the native contacts formed do not generally take similar times to reach the native state, as has been shown for the three-stranded antiparallel β-sheet peptide, Beta3s (**Fig. 1**; ***[29]***). In fact, several order parameters are based on a comparison with a reference structure like the native state (i.e., rmsd, Q, and so on).

Structures that have a native-like values of the order parameters (i.e., high Q, small rmsd, and so on) satisfy a large set of tight constraints on the coordinates of their atoms (i.e., high Q means that a large number of distances between pairs of atoms has to be smaller that a certain tight threshold). As soon as the value of the order parameter becomes less native-like, the number of these constraints decreases or the threshold becomes loose (depending on the order parameter).


Helical
entropic
stabilization
$p_{fold} \approx 0$
TSE
fast relaxation to
the native state
$p_{fold} \approx 0.5$
Native state
$p_{fold} \approx 1$
Curl
enthalpic
stabilization
$p_{fold} \approx 0$
Free-energy
2
1
0
-1
-2
0
0.2
0.4
0.6
0.8
1
Q

This means that a larger and more diverse region of the phase space of the protein projects into the same value of the order parameter. In other words, structures having the same non-native-like order parameter have non-homogeneous structural properties. In **Fig. 1**, some representative conformations of Beta3s with ≈10, ≈50, and ≈80% of the native contacts are shown, from left to right, respectively ***(19)***. A hydrogen bond (HB) is defined as native if the distance between the hydrogen and oxygen atoms is lower than 2.5 Å for more than two-thirds of the conformations belonging to the most populated cluster ***(21)***. A side chain contact (SC) is defined as native if the distance between the center of mass of the two residues averaged over the most populated cluster is smaller than 6.5 Å. Q identifies uniquely one state only when almost all the native contacts are formed, i.e., the native state. For $Q < 70$–80% many heterogenous conformations can have the same number of native contacts ***(19)***.

Most of the time, these conformations are structurally and kinetically heterogeneous (e.g., the $Q \approx 0.5$ conformations with $p_{fold} \approx 0$ and $p_{fold} \approx 0.5$ in **Fig. 1**). In other words, although Q can discriminate between fully folded and fully denatured structures, it does not help in distinguishing structures with properties intermediate between the native and denatured state. Folding times t_{fold} for conformations with half or less of the native contacts formed (central column in **Fig. 1**) can differ as much as two orders of magnitude. Indeed, structures with Q as large as 0.7 may have $t_{fold} \approx 10^2$ ns and, vice versa, Q as low as 0.3 may correspond to structures with $t_{fold} \approx 10^0$ ns.

Of course it can be objected that, in order to optimally describe the thermodynamics and the kinetics of a peptide or a protein, suitable combinations of order parameters can always be found ***(30)***. Even if this possibility exists, it is either very difficult to find and/or very specific for the system under study.

4. The Folding Probability

In the last section, it has been shown that the analysis of the kinetics of a peptide or a protein through near-equilibrium free energy projections can be misleading. Even the energetic barrier between the native and denatured state cannot be reliably estimated from such projections. However, projections are necessary to describe an otherwise very complex system like the one consisting of 10^2–10^5 atoms of a protein. To overcome this problem one has to find a

Fig. 1. (*Opposite page*) Free energy projections on order parameters. In the case of Beta3s ***(21)***, the fraction of native contacts does not necessarily identify structurally and kinetically homogeneous conformations. In the first, second, and third column, conformations with ≈10, ≈50, and ≈80% of native contacts Q are shown, respectively. The projected free energy shows no evidence of the structurally and kinetically heterogeneity of the denatured state of Beta3s (*see* **Subheading 3.**).

projection specifically suited for the wanted features to extract from the simulation, i.e., in the present case, the kinetics of protein folding. The folding probability p_{fold} of a protein conformation saved along a Monte Carlo or MD trajectory is the probability to fold before unfolding ***(14)***. This order parameter defines the kinetics of protein folding because it allows the distinction of structures belonging to the native free energy basin ($p_{fold} = 1$), the unfolded free energy basin ($p_{fold} = 0$), and the free energy barrier ($p_{fold} \approx 0.5$). In principle, it also allows the detection of pathway intermediates in the form of large populations of structures with $0 << p_{fold} << 1$. In other words, it represents the kinetic distance of a structure from the folded state. As in the case of other order parameters, conformations with the same $p_{fold} << 1$ may be structurally different; however, they will have the same kinetic distance from the native state and, in particular, if $p_{fold} \approx 0.5$ they will be unequivocally members of the TSE.

The measure of p_{fold} consists of starting a large number of trajectories from putative TSE structures with varying initial distribution of velocities and counting the number of those that fold within a "commitment" time which has to be chosen much longer than the shortest time scales of conformational fluctuations and much shorter than the average folding time ***(16)***. The concept of p_{fold} calculation originates from a method for determining transmission coefficients, starting from a known transition state ***(31)*** and the identification of simpler transition states in protein dynamics (e.g., tyrosine ring flips) ***(32)***. The approach has been used to identify the otherwise very elusive folding TSE by atomistic Monte Carlo off-lattice simulations of small proteins with a Gō potential ***(16,18)***, as well as implicit solvent MD ***(15,19)*** and Monte Carlo ***(17)*** simulations with a physico-chemical-based potential. The number of trial simulations needed for the reliable evaluation of p_{fold} makes the estimation of the folding probability computationally very expensive. For this reason, we have recently proposed a method to estimate folding probabilities for *all* structures visited in an equilibrium folding-unfolding trajectory without any additional simulation ***(25)***. This method has been applied to the Beta3s peptide and to a large set of its mutants ***(29)***, as will be shown in **Subheading 4.2**.

4.1. Folding Probability of a Single MD Snapshot

For the computation of p_{fold}, a criterion (λ) is needed to determine when the system reaches the folded state. Given a clusterization of the structures, a natural choice for λ is the visit of the most populated cluster, which for structured peptides and proteins is not degenerate (other criteria are also possible, e.g., fraction of native contacts Q larger than a given threshold). Given λ and a commitment time (τ_{commit}), the folding probability $p_{fold}(i)$ of an MD snapshot i is computed as ***(14,16)***:

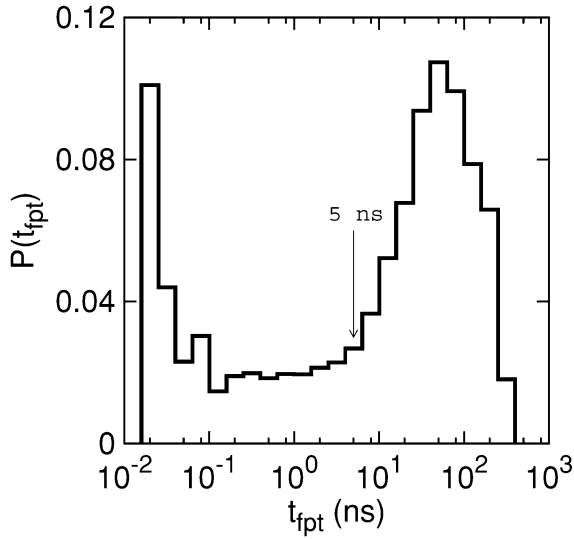


Fig. 2. Probability distribution for the first passage time (fpt) to the most populated cluster (*folded state*) of the DRMS 1.2 Å clusterization of Beta3s.

$$p_{fold}(i) = \frac{n_f(i)}{n_t(i)} \tag{4}$$

where $n_f(i)$ and $n_t(i)$ are the number of trials started from snapshot i, which reach within a time τ_{commit} the folded state and the total number of trials, respectively.

Every simulation started from snapshot i can be considered as a Bernoulli trial of a random variable θ with value 1 (folding within τ_{commit}) or 0 (no folding within τ_{commit}). The variable θ has average and variance on the average of the form:

$$\begin{aligned} \langle\theta\rangle &= p_{fold} = \frac{1}{n_t}\sum_{i=1}^{n_t}\theta_i \\ \sigma^2_{\langle\theta\rangle} &= \frac{1}{n_t}p_{fold}(1-p_{fold}) \end{aligned} \tag{5}$$

where n_t is the total number of trials and the accuracy on the p_{fold} value increases with n_t.

In **Fig. 2** the distribution of the first passage time (fpt) to the folded state of Beta3s is shown. The double peak shape of the distribution provides evidence for the different time scales between intrabasin and interbasin transitions. A value of 5 ns is chosen for τ_{commit} because events with smaller time scales correspond to the diffusion within the native free energy basin, while events with

larger time scales are transitions from other basins to the native one, i.e., folding/unfolding events ***(23)***.

4.2. Folding Probability of a Cluster of Similar Conformations

Conformations that are structurally similar have been shown to have the same kinetic behavior ***(25)***, hence they have similar values of p_{fold}. (Note that the opposite is not necessarily true as already mentioned and as more extensively explained in the next section for the TSE and the denatured state.) Snapshots saved along a trajectory are first grouped in structurally similar clusters. Then, the τ_{commit}-segment of MD trajectory following each snapshot is analyzed to check if the folding condition λ is met (i.e, the snapshot "folds"). For each cluster, the ratio between the snapshots, which lead to folding and the total number of snapshots in the cluster is defined as the cluster-p_{fold} (P_f^C; throughout the text uppercase P and lowercase p refer to folding probability for clusters and individual snapshots, respectively). This value is an approximation of the p_{fold} of any single structure in the cluster which is valid if the cluster consists of structurally similar conformations. In other words, the occurrence of the folding event for the snapshots of a given cluster can be considered as a Bernoulli trial of a random variable θ. The average of θ and variance on the average for the set of snapshots belonging to a given cluster α can be written as:

$$\begin{aligned} P_f^C[\alpha] = \langle\theta\rangle &= \frac{1}{W}\sum_{i=1}^{W}\theta_i, \quad i \in \alpha \\ \sigma^2_{\langle\theta\rangle} &= \frac{1}{W}P_f^C(1-P_f^C) \end{aligned} \tag{6}$$

where W is the number of snapshots in cluster α. P_f^C is the average folding probability over a set of structurally homogeneous conformations. Using the clustering and the folding criterion λ introduced previously, values of P_f^C for the 78,183 clusters of Beta3s can be computed by **Eq. 6**, i.e., the number of conformations of the cluster that fold within 5 ns divided by the total number of conformations belonging to the cluster.

$$P_f[\alpha] = \frac{1}{W}\sum_{i=1}^{W} p_{fold}(i), \quad i \in \alpha \tag{7}$$

which is measured by starting several simulations from each snapshot i in the cluster α with W snapshots, is well approximated by P_f^C whose evaluation is straightforward.

To compare the values of P_f^C with those obtained from the standard approach ***(14)***, folding probabilities P_f were computed for the structures of 37 clusters by starting several 5-ns MD runs from each structure and counting those that fold

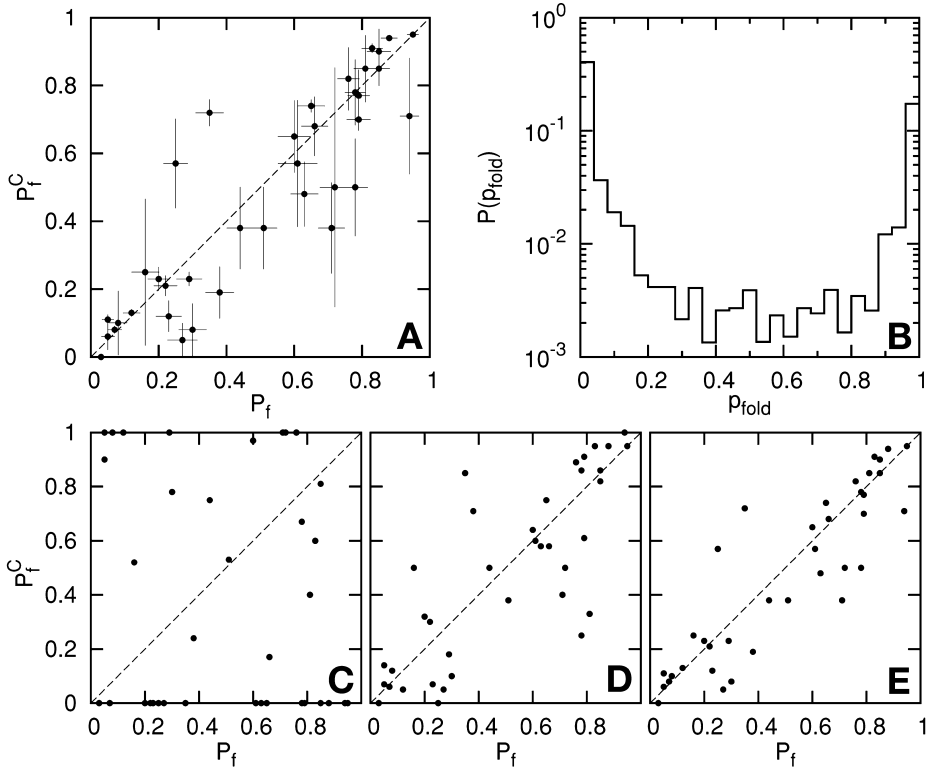


Fig. 3. Cluster folding probability P_f^C. **(A)** Scatter plot of P_f^C vs P_f. The DRMS 1.2 Å clusterization and the folding criterion λ (reaching the most populated cluster within τ_{commit} = 5 ns) were used. **(B)** Probability distribution of the p_{fold} value for the 500,000 snapshots saved along the 10-μs MD trajectory of Beta3s. The folding probability for snapshot i is computed as $p_{fold}(i) = P_f^C[\alpha]$ for $i \in \alpha$. **(C–E)** Scatter plot of P_f^C vs P_f for 1.0, 5.0, and 10 μs of simulation time, respectively.

(**Eqs. 4** and **7**). The 37 clusters chosen among the 78,183 include both high- and low-populated clusters with P_f^C values evenly distributed in the range between 0 and 1. In the case of large clusters, a subset of snapshots is considered for the computation of P_f. In those cases W is replaced in **Eq. 7** by $W_{sample} < W$ that is the number of snapshots involved in the calculation. Namely, for the 37 clusters previously mentioned, a correlation of 0.89 between P_f^C and P_f is found with a slope of 0.86 (*see* **Fig. 3A**), indicating that the procedure is able to estimate folding probabilities for clusters on the folding–transition barrier ($P_f \approx 0.5$) as well as in the folding ($P_f \approx 1.0$) or unfolding ($P_f \approx 0.0$) regions. The error bars for P_f^C in **Fig. 3A** are derived from the definition of variance given in **Eq. 6**. In the same spirit of **Eq. 6** the folding probability P_f and its variance are written as:

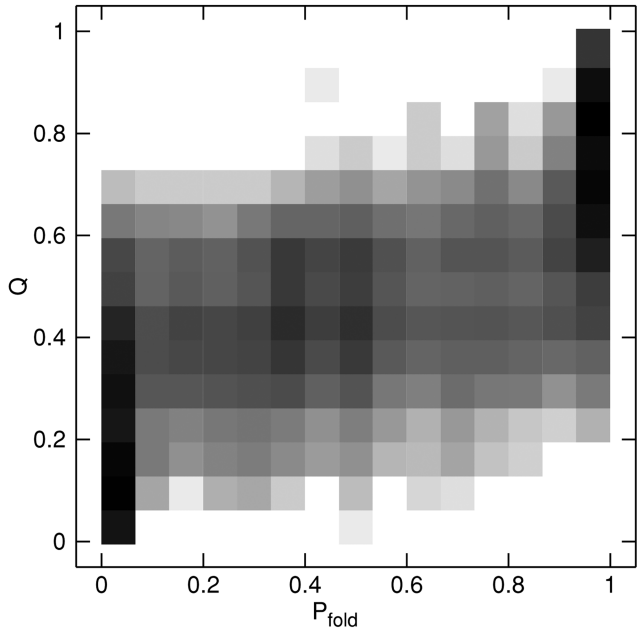


Fig. 4. Distribution of fraction of native contacts Q and P_f^C in the wild-type Beta3s simulations. The gray scale from black to white corresponds to high and low density, respectively. Although structures with very large Q ($Q > 0.8$) or very low Q ($Q < 0.2$) have P_f^C close to 1 or 0, respectively, conformations with intermediate values of Q span all the allowed spectrum of P_f^C values.

$$P_f = \langle\theta\rangle = \frac{1}{N}\sum_{i=1}^{N}\theta_i$$
$$\sigma^2_{\langle\theta\rangle} = \frac{1}{N}P_f(1-P_f) \tag{8}$$

where $N = \sum n_t$ is the total number of runs and θ is equal to 1 or 0, if the run folded or unfolded, respectively. Note that the same number of runs n_t has been used for every snapshot of a cluster. The large vertical error bars in **Fig. 3A** correspond to clusters with less than 10 snapshots. The largest deviations between P_f and P_f^C are around the 0.5 region. This is owing to the limited number of crossings of the folding barrier observed in the MD simulation (**Fig. 3B**, around 70 events of folding ***[23]***). Improvements in the accuracy for the estimation of P_f are achieved as the number of folding events, i.e., the simulation time, increases (**Fig. 3C–E**).

The validity of P_f^C as an approximation of P_f, is robust with respect to the choice of the clusterization. Similar results can be obtained also with different flavors of conformation space partitioning, as long as they group together structurally

homogeneous conformations, e.g., clusterization based on root mean square deviation of atomic coordinates (RMSD) or secondary structure strings ***(25)***. The latter are appropriate for structured peptides but not for proteins with irregular secondary structure because of string degeneracy. Note that partitions based on order parameters (like native contacts) are usually unsatisfactory and not robust. This is mainly owing to the fact that clusters defined in this way are characterized by large structural heterogeneities ***(19)***.

Interestingly, there is no correspondence between the number of native contacts formed and p_{fold} (**Fig. 4**). In other words, it would have been impossible to simply use the order parameter Q to extract TSE conformations. This result shows again that the indiscriminate use of free energy projections on order parameters can be misleading and kinetic properties cannot, in general, be inferred from the thermodynamic analysis.

5. The Transition State Ensemble Defined Using the Folding Probability

The folding probability of structure i is estimated as $p_{fold}(i) = P_f^C[\alpha]$ for $i \in \alpha$. This approximation allows one to plot the pairwise RMSD distribution of Beta3s structures with $p_{fold} > 0.51$ (native state), $0.49 < p_{fold} < 0.51$ (TSE), and $p_{fold} < 0.49$ (denatured state) (**Fig. 5A**). For the native state, the distribution is peaked around low values of RMSD (≈1.5 Å) indicating that structures with $p_{fold} > 0.51$ are structurally similar and belong to a nondegenerate state. The statistical weight of this group of structures is 49.4% and corresponds to the expected statistics for the native state because the simulations are performed at the melting temperature. In the case of TSE, the distribution is broad because of the coexistence of heterogeneous structures. This scenario is compatible with the presence of multiple folding pathways. Beta3s folding was already shown to involve two main average pathways depending on the sequence of formation of the two hairpins ***(19,21)***. Here, a *naive* approach based on the number of native contacts ***(21,25)*** is used to structurally characterize the folding barrier. TSE structures with number of native contacts of the first hairpin greater than the ones of the second hairpin are called type I conformations (**Fig. 5B**), otherwise they are called type II (**Fig. 5C**). In both cases the transition state is characterized by the presence of one of the two native hairpins formed while the rest of the peptide is mainly unstructured. These findings are also in agreement with the complex network analysis of Beta3s reported recently ***(19)***. Finally, the denatured state shows a broad pairwise RMSD distribution around even larger values of RMSD (≈5.5 Å), indicating the presence of highly heterogeneous conformations.

6. Φ-Value Analysis by MD Simulations

The p_{fold} values of all conformations saved along a reversible folding MD trajectory can be used to isolate the TSE. However, the information on TSE derived

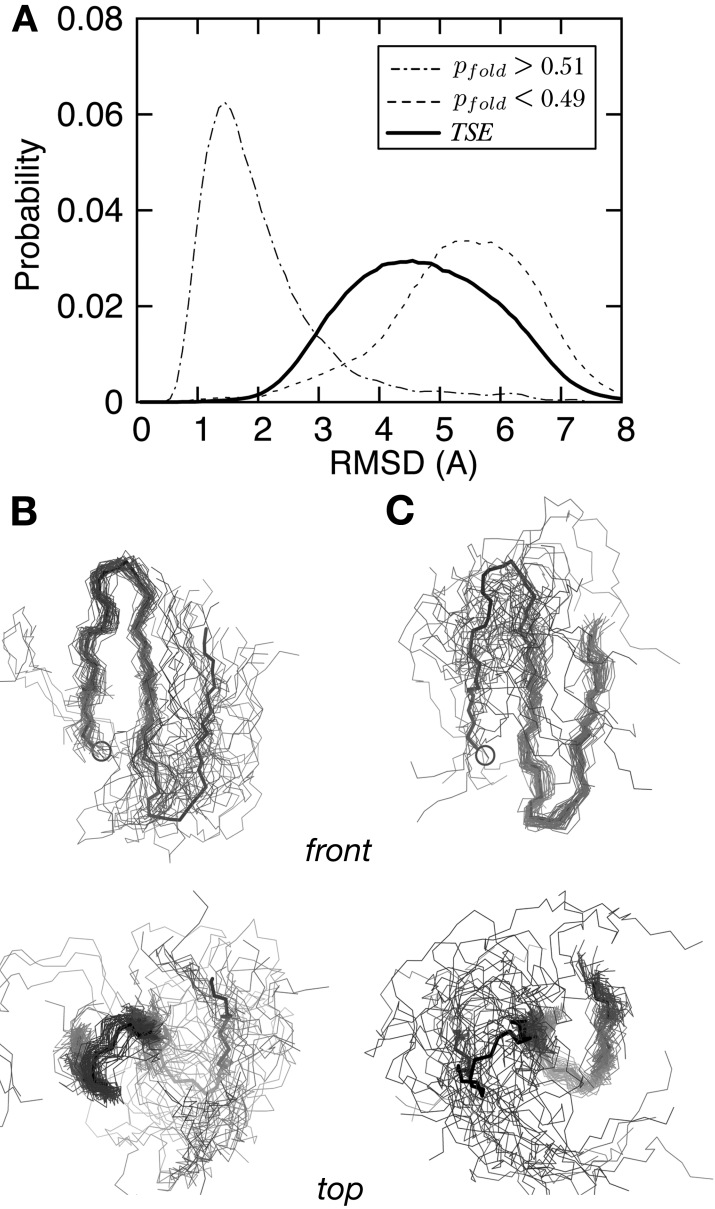


Fig. 5. Transition state ensemble (TSE) of Beta3s. **(A)** Distribution of the values of pairwise RMSD for structures with $p_{fold} > 0.51$ (native state), $0.49 < p_{fold} < 0.51$ (TSE), and $p_{fold} < 0.49$ (denatured state). **(B)** Type I and **(C)** type II transition states (thin lines). Structures are superimposed on residues 2–11 and 10–19 with an average pairwise RMSD of 0.81 and 0.82 Å for type I and type II, respectively. For comparison, the native state is shown as a thick line with a circle to label the N-terminus.

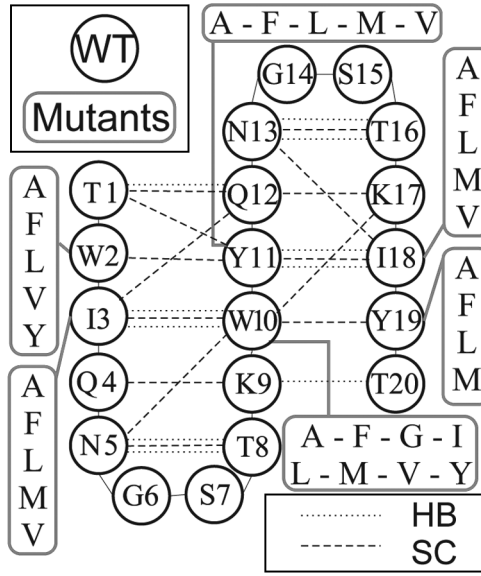


Fig. 6. Schematic representation of the Beta3s peptide, where the wild-type sequence and the mutants are indicated. The backbone hydrogen bonds (dotted lines) and side chain contacts (dashed lines) common to most of the peptides are reported. HB, hydrogen bond; SC, side chain contacts; WT, wild-type.

from protein folding experiments is represented by the Φ values. As we will see in more detail, the Φ value of a residue is the change in the activation free energy of folding relative to the change in stability of the protein on mutation of the residue. Φ values have been usually interpreted as the fraction of native contacts formed at TSE by the mutated residue. This interpretation, however, does not allow one to consider non-native interactions that may form at TSE and is not able to explain anomalous Φ values (i.e., those out of the 0 to 1 range). Thus, we have extensively tested the standard interpretation of Φ values by evaluating them from the folding and unfolding rates measured in equilibrium MD simulations of wild-type Beta3s and a large number of single-point mutants.

Thirty-two single-point mutations of the hydrophobic and aromatic side chains W2, I3, W10, Y11, I18, and Y19 were investigated (**Fig. 6**). The six sites of mutation are distributed along the sequence of the peptide, two for each strand. Between four and eight mutations have been studied for each site. Six of the 32 mutations are nondisruptive (I3A, I3V, Y11F, I18A, I18V, and Y19F), six mutations are conservative but change the steric properties of the side chain (I3M, Y11L, Y11M, I18M, Y19L, and Y19M), and the remaining 20 mutations are radical but acceptable because, in most of the cases, they do not significantly change the TSE of the peptide. This is probably because of the fact that

the side chains of Beta3s are not fully buried in a densely packed hydrophobic core as is the case in larger proteins ***(33)***. Ten MD runs of 2 μs each (total of 20 μs for each mutant) with different initial velocities were performed with the Berendsen thermostat at 330 K, which is close to the melting temperature of wild-type Beta3s ***(34)***. A time step of 2 fs was used and the coordinates were saved every 20 ps for a total of 10^6 conformations for each mutant. During the 20-μs simulation time between 57 and 120 folding events were observed for every mutant (**Table 1**), thus providing sufficient statistical sampling for the kinetic analysis. The small statistical error is supported by the small difference in the native population measured for each individual mutant on two disjoint equal-size subsets of the trajectories (5% on average, the largest being 13%).

The native structure of the wild-type, i.e., the three-stranded anti-parallel β-sheet with turns at G6-S7 and G14-S15, is also the most populated in all the mutants, as shown by the cluster analysis of the trajectories (**Table 1**). The only exception is Y11V, which has a more distorted native state and has not been considered for further analysis. Moreover, there is no predominant structure in the denatured state for any of the mutants. The number of folding and unfolding events observed along the trajectories ranges from 57 to 120 and 64 to 127, respectively (**Table 1**). Interestingly, the values of the stability change upon mutation, calculated with **Eq. 2**, and show that all mutants are less stable than wild-type Beta3s except for W10F and I3V, which are essentially as stable as Beta3s. This result is not unexpected because Beta3s is a designed peptide whose sequence was carefully optimized for its fold ***(20)***.

As in the kinetic experiments used to measure experimental Φ values, free energy changes with respect to wild-type are computed from the folding and unfolding rates. The fraction of native contacts Q has been computed along the trajectories of all peptides. A folding (unfolding) event occurs when, along the trajectory, Q first reaches values larger than 0.85 (lower than 0.15) immediately after a previous unfolding (folding) event ***(21)***. All the trajectories are started from the folded state, thus, the first event is always an unfolding. The average time separation between a folding (unfolding) event and the previous unfolding (folding) event, is the folding (unfolding) time τ_f (τ_u). The folding and unfolding rates are $k_f = 1/\tau_f$ and $k_u = 1/\tau_u$, respectively. Setting the free energy of the denatured state as reference:

$$\Delta\Delta G^{kin}_{TS-D} = RT \log\left(\frac{k_f^{WT}}{k_f^{mut}}\right) \tag{9}$$

$$\Delta\Delta G^{kin}_{N-D} = RT \log\left(\frac{k_f^{WT}}{k_f^{mut}} \cdot \frac{k_u^{mut}}{k_u^{WT}}\right) \tag{10}$$

Table 1
Stability, Folding/Unfolding Rates, and Φ Values of the Mutants

Mutation [a]	W_{highQ} [b] (%)	*Nat.Cont.* [c]	W_{lowQ} [d] (%)	τ_f [e] (ns)	N_f [f]	τ_u [g] (ns)	N_u [h]	$\Delta\Delta G^{kin}_{N-D}$ [i] (kcal/mol)	$\Delta\Delta G^{kin}_{TS-D}$ [i] (kcal/mol)	Φ [i,j]
WT	21.4	19.3 ± 1.7	2.9	70 ± 10	92	67 ± 6	94			
W2A	26.5	18.1 ± 2.3	3.5	107 ± 14	108	63 ± 6	114	-0.32 ± 0.15	-0.28 ± 0.13	0.87 ± 0.57
W2F	33.5	18.8 ± 2.2	3.4	106 ± 14	97	82 ± 8	103	-0.14 ± 0.16	-0.27 ± 0.13	-
W2L	24.9	18.2 ± 2.2	6.3	109 ± 16	101	63 ± 5	111	-0.34 ± 0.16	-0.30 ± 0.14	0.87 ± 0.57
W2V	23.6	18.3 ± 2.3	4.4	124 ± 17	95	62 ± 6	102	-0.43 ± 0.16	-0.38 ± 0.13	0.89 ± 0.45
W2Y	21.9	18.5 ± 2.4	6.4	129 ± 21	93	65 ± 6	98	-0.43 ± 0.16	-0.41 ± 0.14	0.95 ± 0.49
I3A	19.9	18.7 ± 2.2	3.9	137 ± 18	92	64 ± 5	101	-0.48 ± 0.15	-0.44 ± 0.13	0.93 ± 0.40
I3F	33.0	18.8 ± 2.1	3.3	121 ± 22	83	93 ± 8	91	-0.15 ± 0.17	-0.36 ± 0.15	-
I3L	28.5	18.5 ± 2.4	3.9	119 ± 19	94	72 ± 7	101	-0.31 ± 0.17	-0.35 ± 0.14	1.1 ± 0.77
I3M	30.2	18.9 ± 2.2	5.4	108 ± 19	94	81 ± 9	102	-0.16 ± 0.17	-0.29 ± 0.15	-
I3V	37.2	18.6 ± 2.1	5.2	124 ± 18	75	109 ± 10	83	-0.06 ± 0.16	-0.38 ± 0.14	-
W10A	31.8	19.5 ± 2.1	5.0	161 ± 21	74	95 ± 10	79	-0.32 ± 0.16	-0.55 ± 0.13	1.7 ± 0.93
W10F	41.3	18.7 ± 2.2	3.8	77 ± 9	120	78 ± 6	127	0.04 ± 0.14	-0.06 ± 0.12	-
W10G	12.7	19.3 ± 2.2	3.1	212 ± 32	60	68 ± 9	69	**-0.72**± 0.17	-0.73 ± 0.14	**1.0 ± 0.31**
W10I	30.8	18.3 ± 2.1	6.0	129 ± 17	77	88 ± 9	83	-0.23 ± 0.16	-0.40 ± 0.13	-
W10L	20.8	18.8 ± 2.2	4.2	166 ± 22	81	58 ± 5	87	**-0.67**± 0.16	-0.57 ± 0.13	**0.86 ± 0.28**
W10M	18.4	19.0 ± 2.2	6.6	155 ± 21	82	52 ± 5	91	**-0.68**± 0.16	-0.52 ± 0.13	**0.76 ± 0.26**
W10V	17.2	17.8 ± 2.5	6.7	259 ± 40	57	65 ± 11	64	**-0.88**± 0.19	-0.86 ± 0.14	**0.98 ± 0.26**
W10Y	26.2	19.0 ± 2.1	3.5	118 ± 15	94	77 ± 7	98	-0.26 ± 0.15	-0.35 ± 0.13	-
Y11A	5.7	18.1 ± 2.0	2.3	249 ± 38	64	30 ± 3	71	**-1.37**± 0.17	-0.84 ± 0.14	**0.61 ± 0.13**
Y11F	33.1	19.1 ± 2.2	4.4	138 ± 20	73	112 ± 12	79	-0.11 ± 0.16	-0.45 ± 0.14	-
Y11L	14.8	18.6 ± 2.1	4.8	169 ± 23	76	54 ± 6	83	**-0.72**± 0.16	-0.58 ± 0.13	**0.81 ± 0.26**
Y11M	11.3	18.0 ± 2.2	3.5	152 ± 24	95	35 ± 3	105	**-0.94**± 0.16	-0.51 ± 0.14	**0.54 ± 0.18**
Y11V	5.7	17.0 ± 2.7	7.4							
I18A	12.3	18.5 ± 2.3	2.4	168 ± 22	80	53 ± 6	88	**-0.73**± 0.16	-0.58 ± 0.13	**0.79 ± 0.25**
I18F	21.3	19.0 ± 2.0	3.2	159 ± 23	74	72 ± 8	83	-0.50 ± 0.17	-0.54 ± 0.14	1.1 ± 0.46
I18L	22.2	19.0 ± 2.2	4.4	145 ± 19	73	94 ± 9	81	-0.26 ± 0.16	-0.48 ± 0.13	-
I18M	28.9	18.8 ± 2.2	4.8	97 ± 15	99	77 ± 6	106	-0.13 ± 0.16	-0.22 ± 0.14	-
I18V	29.6	18.8 ± 2.3	3.2	124 ± 20	87	86 ± 9	93	-0.22 ± 0.17	-0.38 ± 0.14	-
Y19A	20.7	18.6 ± 2.4	7.4	123 ± 18	90	84 ± 8	95	-0.23 ± 0.16	-0.37 ± 0.14	-
Y19F	29.2	18.4 ± 2.2	3.8	130 ± 18	92	71 ± 7	98	-0.37 ± 0.16	-0.41 ± 0.13	1.1 ± 0.59
Y19L	30.0	18.3 ± 2.2	3.2	117 ± 17	83	88 ± 8	89	-0.17 ± 0.16	-0.34 ± 0.13	-
Y19M	17.5	18.5 ± 2.3	6.2	155 ± 26	68	97 ± 10	76	-0.28 ± 0.17	-0.52 ± 0.15	-

[a] Mutants in italics are radical but acceptable and mutations in Roman are conservative.
[b] Statistical weight of the three most populated clusters with $Q \geq 16/24$.
[c] Average number of contacts in the three most populated clusters with $Q \geq 16/24$.
[d] Statistical weight of the three most populated with $Q < 16/24$.
[e] Average folding time.
[f] Number of folding events.
[g] Average unfolding time.
[h] Number of unfolding events.
[i] The standard deviations have been obtained by propagation of the error on τ_f and τ_u.
[j] Dashes indicate "unreliable" Φ values due to $|\Delta\Delta G^{kin}_{N-D}|$ <0.3 kcal/mol. Boldface emphasize the "reliable" Φ values and the corresponding large stability changes (33). The multipoint Φ values are 0.77, 0.60, 0.79, 0.46, 0.72, and 1.23 for W2, I3, W10, Y11, I18, and Y19, respectively.

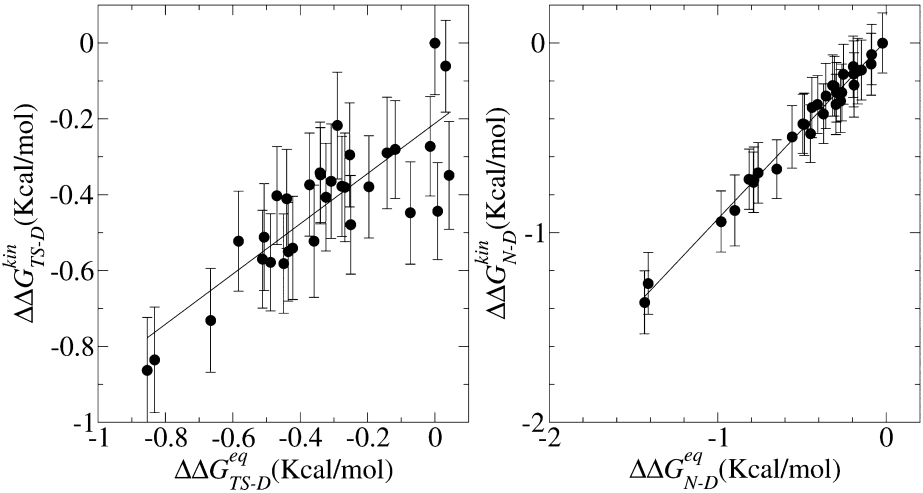


Fig. 7. Comparison between free energy changes calculated with the kinetic and P_f^C data. The correlation coefficient is 0.83 and 0.99 for $\Delta\Delta G_{TS-D}$ and $\Delta\Delta G_{N-D}$, respectively.

The Φ value is $\Phi = \Delta\Delta G^{kin}_{TS-D}/\Delta\Delta G^{kin}_{N-D}$. Values of $\Delta\Delta G^{kin}_{TS-D}$ and $\Delta\Delta G^{kin}_{N-D}$ from multiple mutations at the same site can be displayed on a single plot. The slope of the corresponding regression line is called the multipoint Φ value ***(33,35)***.

Clusters are assigned to the native state, the TSE and the denatured state assemble according to their P_f^C. Their statistical weights are W_N, W_{TS} and W_D, respectively; these values can be used to evaluate relative free energies by a different equation with respect to the kinetically evaluated $\Delta\Delta G^{kin}$. In the canonical ensemble $\Delta G^{eq}_{TS-D} = -RT \log(W_{TS}/W_D)$ and $\Delta G^{eq}_{N-D} = -RT \log(W_N/W_D)$. As shown in **Fig. 7**, an excellent match is observed between the $\Delta\Delta G^{kin}_{N-D}$ and $\Delta\Delta G^{eq}_{N-D}$ values (correlation coefficient of 0.99) and a good correlation between $\Delta\Delta G^{kin}_{TS-D}$ and $\Delta\Delta G^{eq}_{TS-D}$ (correlation coefficient of 0.83). The agreement represents a consistency check for the parameters used to define folding and unfolding events. That activation free energy differences computed with the two sets of data show larger discrepancies than changes in stability, is owing to the difficulty in sampling the TSE.

6.1. Accuracy of Two-Point and Multipoint Φ Values

Figure 8 shows the Φ values extracted from the simulations as a function of the change in free energy of folding upon mutation (*see also* **Table 1**). Because of the difficulties in the interpretation of Φ values, as many mutants as possible have been considered and the resulting Φ values divided into classes of "reliable", "tolerable", and "unreliable" according to the size of the induced stability change

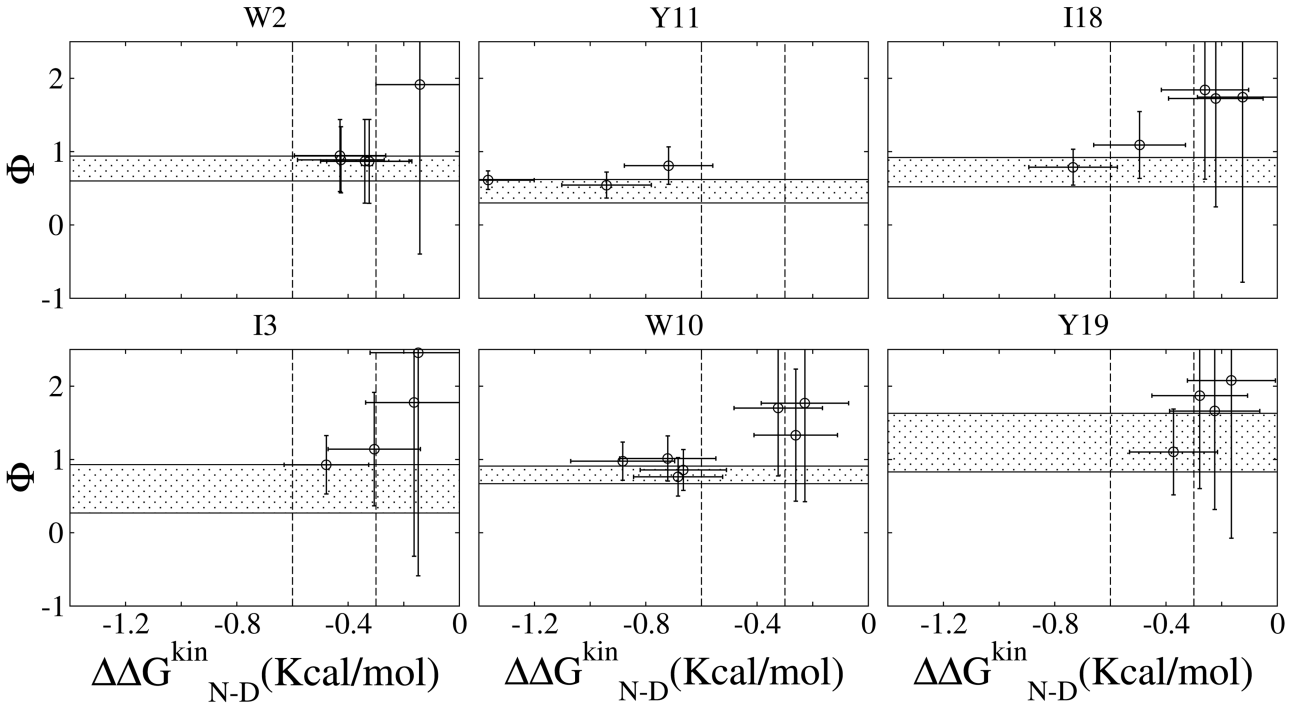


Fig. 8. Φ values as a function of change in the native state stability upon mutation. (The shadowed horizontal region indicates one standard deviation around the multipoint Φ value. The Φ values span a wide range and become "anomalous" for $|\Delta\Delta G^{kin}_{N-D}|$ smaller than about 0.3 kcal/mol. The Φ values corresponding to mutations with $|\Delta\Delta G^{kin}_{N-D}| > 0.3$ are mainly in the "normal" range, i.e., between 0 and 1, and are in agreement with the multipoint Φ value. Vertical dashed lines are drawn at $\Delta\Delta G^{kin}_{N-D} = -0.3$ kcal/mol and $\Delta\Delta G^{kin}_{N-D} = -0.6$ kcal/mol. The Φ value of mutations I3V, W10F, and Y11F are located outside of the plot boundaries.) The graphs are ordered according to the antiparallel β-sheet topology of Beta3s with vertical orientation of the three strands, and the N- and C-terminus on the top-left and bottom-right, respectively.

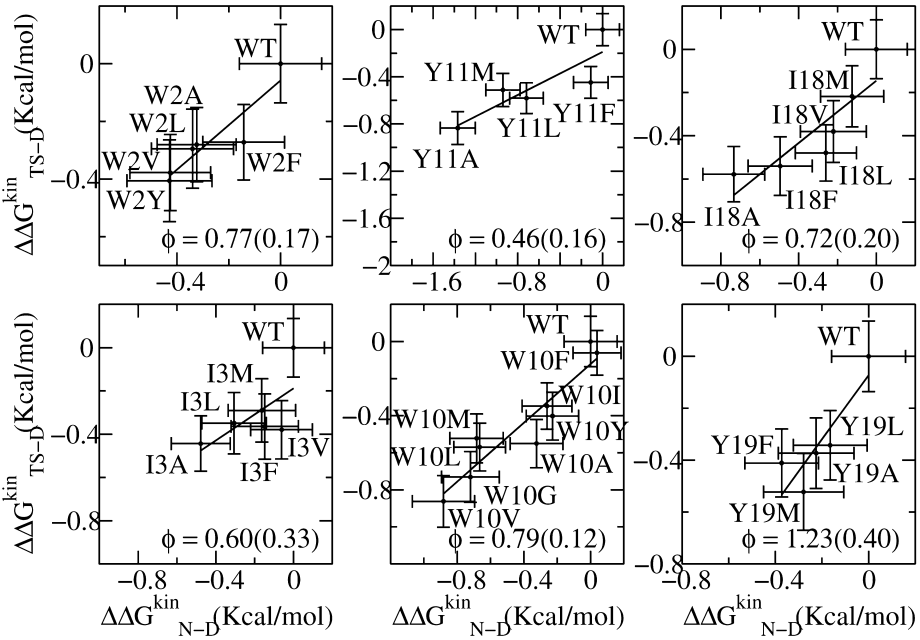


Fig. 9. $\Delta\Delta G^{kin}_{TS-D}$ plotted vs $\Delta\Delta G^{kin}_{N-D}$ for all the mutants grouped according to the mutation site along the structure of Beta3s. The optimal regression line (including the wild-type data point) is plotted and its slope, i.e., the multipoint Φ value, is reported in the lower-right corner of each graph with the standard deviation derived from the fit in parentheses. The correlation coefficient is 0.91, 0.67, 0.93, 0.86, 0.87, and 0.88 for W2, I3, W10, Y11, I18, and Y19 mutants, respectively.

$|\Delta\Delta G^{kin}_{N-D}|$. The deviations from the 0 to 1 range are large for "unreliable" Φ values, i.e., for mutations with $|\Delta\Delta G^{kin}_{N-D}|<0.3$, in agreement with previous observations ***(35)***. Indeed, in the "unreliable" class the deviation can be observed for both radical mutations (e.g., I3F, W10A, Y19A) as well as for nondisruptive mutations (e.g., I3V, Y11F, and I18V). For "tolerable" Φ values, i.e., $0.3 \le |\Delta\Delta G^{kin}_{N-D}| < 0.6$, the deviation from the 0–1 interval is less frequent but the relative error is large. The eight "reliable" Φ values ($|\Delta\Delta G^{kin}_{N-D}|\ge 0.6$) are all in the range 0 to 1 and have a small standard deviation. In a small-structured peptide, like Beta3s, most residues have a relatively large exposed surface area in the folded state so that conservative mutations generally induce small free energy changes. Indeed, among the six conservative mutations only I18A falls in the "reliable" class. For this reason more radical mutations have been also investigated.

The multipoint Φ of Beta3s as extracted from the simulations are reported in **Fig. 9**. The good linear relationship between $\Delta\Delta G^{kin}_{TS-D}$ and $\Delta\Delta G^{kin}_{N-D}$, observed in mutants of W2, W10, Y11, and Y19, supports the validity of the multipoint

analysis for these residues and indicates a substantial similarity among the folding TSEs of those peptides. In mutants of I3 the linear correlation is less strong than the others and in I18 there is a change in the slope for $\Delta\Delta G^{kin}_{N-D} < -0.3$ kcal/mol. A possible explanation for the presence of a linear relationship in the multipoint plots is the partial flexibility of the native state of Beta3s ***(19)***. Its partially exposed non-polar side chains, which have been mutated in this work, are involved in less specific interactions with the rest of the peptide than buried side chains in the hydrophobic core of larger proteins. Because of the partial flexibility, the mutations do not affect only specific interactions but produce an effect that is spread over the large available set of contacts and thus averaged over them. This averaging of the effects of mutations in the native state may translate into a simple linear dependence of the effects in the TS. In this context, deviations from linearity may indicate TSE shifts.

In multipoint plots different local probes of the same residue are forced in a single fit which can yield wrong estimates ***(36)***. As an example, in the I → V → A → G mutation series the I → V measures interactions originating from tertiary structure contacts, the V → A a mixture of tertiary and secondary structure interactions, whereas the A → G reports almost exclusively on secondary structure formation ***(36)***. In a framework ***(37)*** or diffusion–collision ***(38)*** mechanism of folding, the "tertiary" Φ values will most probably be lower than "secondary" Φ values, even for the same residue. In the case of Beta3s, where the formation of β-sheet backbone hydrogen bonds and long-range contacts between side chains are concomitant events (*see* **Fig. 4** in **ref. *21***), different mutations probe the formation of the same level of structure (i.e., the β-sheet) with no distinction between secondary and tertiary components. This supports the validity of the multipoint analysis for Beta3s, which we do not want to generalize to proteins with more complex folds.

Given the peculiarities of Beta3s, i.e., concomitant formation of secondary and tertiary structure and partial flexibility of its folded state, multipoint Φ values may add information on the accuracy of the two-point Φ values. Indeed, "reliable" and "tolerable" Φ values fall mostly within a standard deviation from the corresponding multipoint Φ value (**Fig. 8**), whereas "unreliable" Φ values show large deviations. Five of the six multipoint Φ values of Beta3s are larger than 0.5. For diffuse TSEs of proteins of about 100 residues, Φ values around 0.2–0.3 have been measured experimentally ***(39,40)***. The high Φ values of Beta3s are probably owing to the small size of the peptide. Because of its small size a large part of the native interactions of the hydrophobic residues is already present in the rate-limiting step.

6.2. Structural Interpretation of Φ Values

In each snapshot a van der Waals contact is defined when the distance between two heavy atoms is smaller than 6 Å. $p_N(i)$ and $p_{TS}(i)$ measure the fraction of native

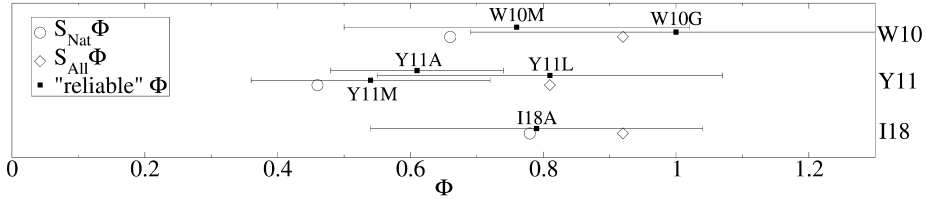


Fig. 10. Comparison between "reliable" two-point Φ values (filled squares) of mutants with TSE similar to wild-type, and the structure of wild-type TSE as measured by $S_x\Phi$ values (open symbols). The structural Φ values are the ratio between the number of contacts formed in TSE and native state. $S_{Nat}\Phi$ takes into account only native contacts, whereas $S_{All}\Phi$ includes native and non-native contacts. The two-point Φ values tend to overestimate the degree of nativeness of the TSE (measured by $S_{Nat}\Phi$) because of the presence of specific non-native interactions.

and TSE structures, respectively, in which the contact i is formed. If $p_N(i) > 0.66$ the contact i belongs to the set of the native contacts (*NC*). The structural Φ value:

$$S_{Nat}\Phi(R) = \frac{1}{M_{NC(R)}} \frac{\sum_{i \in NC(R)} p_{TS}(i)}{\sum_{i \in NC(R)} p_N(i)}, \tag{11}$$

where $M_{NC}(R)$ is the number of native contacts of residue *R*, represents an estimate of the degree of nativeness of residue *R* at the TSE. This measure has been used in the past to give a structural interpretation to experimental Φ values ***(41–43)***. An estimate of the relevance of non-native interactions at the TSE is obtained by extending the sum to all possible contacts (*AC*), including contacts not present in the *NC* set:

$$S_{All}\Phi(R) = \frac{1}{M_{AC(R)}} \frac{\sum_{i \in AC(R)} p_{TS}(i)}{\sum_{i \in AC(R)} p_N(i)} \tag{12}$$

Both $S_{Nat}\Phi$ and $S_{All}\Phi$ profiles of Beta3s provide a detailed picture of its TSE. It is useful to compare them with the "reliable" Φ values, i.e., those derived from mutations that do not significantly change the TSE of the peptide (e.g., W10M, W10G, Y11A, Y11L, Y11M, and I18A). Such comparison allows for the assessment of the standard interpretation of the Φ as the ratio between contacts formed at TSE and native state (**Fig. 10**). The comparison reveals that, within their error, the two-point Φ values are in agreement with both $S_x\Phi$. However, the former tend to overestimate the degree of native structure present at the TSE (i.e., "reliable" $\Phi > S_{Nat}\Phi$) because specific non-native interactions

are formed at the TSE ***(29)***. More generally speaking, the presence of specific non-native contacts, distinguishing the TSE conformations from other structures having the same native interactions but different non-native interactions, makes the standard interpretation of Φ values not completely appropriate. Namely, neglecting non-native interactions may prevent a complete understanding of the factors that are responsible for protein folding.

7. Conclusions

Despite its very simple native topology, the 20-residue structured peptide Beta3s has been shown, using MD simulations with implicit solvent, to have multiple folding pathways ***(21)*** and a very heterogeneous denatured state consisting of both high enthalpy, high entropy basins, and low enthalpy, low entropy traps ***(19)***. Furthermore, folding-unfolding equilibrium simulations of Beta3s and several single-point mutants have been used to evaluate folding probabilities of Beta3s conformations ***(25)*** and Φ values of several of its residues ***(29)***, respectively. The latter, calculated from folding and unfolding rates measured from the MD trajectories, are reliable if the stability loss upon mutation is larger than about 0.6 kcal/mol, in agreement with experimental observations. Another interesting simulation result is that Φ values tend to overestimate the nativeness of the TSE, when interpreted neglecting the non-native interactions. The next challenge is to generalize the simulation results obtained with Beta3s to other structured peptides and small proteins.

References

1. Abkevich, V., Gutin, A., and Shakhnovich, E. (1994) Free-energy landscape for protein-folding kinetics: intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101,** 6052–6062.
2. Dill, K. and Chan, H. (1997) From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4,** 10–19.
3. Frauenfelder, H., Sligar, S., and Wolynes, P. (1991) The energy landscapes and motions of proteins. *Science* **254,** 1598–1603.
4. Saven, J., Wang, J., and Wolynes, P. (1994) Kinetics of protein-folding: the dynamics of globally connected rough energy landscapes with biases. *J. Chem. Phys.* **101,** 11037–11043.
5. Wang, J., Onuchic, J., and Wolynes, P. G. (1996) Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Phys. Rev. Lett.* **76,** 4861–4864.
6. Daggett, V. and Fersht, A. (2003) Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* **28,** 18–25.
7. Leopold, P. E., Montal, M., and Onuchic, J. N. (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA* **89,** 8721–8725.

8. Karplus, M. (1997) The Levinthal paradox: yesterday and today. *Fold. Des.* **2,** S69–S75.
9. Fersht, A. R. (1999) *Structure and Mechanism in Protein Science: Guide to Enzyme Catalysis and Protein Folding,* W. H. Freeman, New York, NY.
10. Fersht, A. R., Matouschek, A., and Serrano, L. (1992) The folding of an enzyme. 1. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224,** 771–782.
11. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S., and Karplus, M. (1983) Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4,** 187–217.
12. Karplus, M. and Kuriyan, J. (2005) Chemical theory and computation special feature: molecular dynamics and protein function. *Proc. Natl. Acad. Sci. USA* **102,** 6679–6685.
13. McCammon, J. A., Gelin, B. R., and Karplus, M. (1977) Dynamics of folded proteins. *Nature* **267,** 585–590.
14. Du, R., Pande, V., Grosberg, A., Tanaka, T., and Shakhnovich, E. (1998) On the transition coordinate for protein folding. *J. Chem. Phys.* **108,** 334–350.
15. Gsponer, J. and Caflisch, A. (2002) Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci. USA* **99,** 6719–6724.
16. Hubner, I. A., Shimada, J., and Shakhnovich, E. I. (2004) Commitment and nucleation in the protein G transition state. *J. Mol. Biol.* **336,** 745–761.
17. Lenz, P., Zagrovic, B., Shapiro, J., and Pande, V. S. (2004) Folding probabilities: a novel approach to folding transitions and the two-dimensional Ising-model. *J. Chem. Phys.* **120,** 6769–6778.
18. Li, L. and Shakhnovich, E. I. (2001) Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. USA* **98,** 13014–13018.
19. Rao, F. and Caflisch, A. (2004) The protein folding network. *J. Mol. Biol.* **342,** 299–306.
20. De Alba, E., Santoro, J., Rico, M., and Jimenez, M. (1999) De novo design of a monomeric three-stranded antiparallel beta-sheet. *Protein Sci.* **8,** 854–865.
21. Ferrara, P. and Caflisch, A. (2000) Folding simulations of a three-stranded antiparallel beta-sheet peptide. *Proc. Natl. Acad. Sci. USA* **97,** 10780–10785.
22. Ferrara, P., Apostolakis, J., and Caflisch, A. (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* **46,** 24–33.
23. Cavalli, A., Haberthur, U., Paci, E., and Caflisch, A. (2003) Fast protein folding on downhill energy landscape. *Protein Sci.* **12,** 1801–1803.
24. Hartigan, J. (1975) *Clustering Algorithms,* Wiley, New York, NY.
25. Rao, F., Settanni, G., Guarnera, E., and Caflisch, A. (2005) Estimation of protein folding probability from equilibrium simulations. *J. Chem. Phys.* **122,** 184901.
26. Chan, H. S. and Dill, K. A. (1998) Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins* **30,** 2–33.
27. Pande, V., Grosberg, A., Tanaka, T., and Rokhsar, D. (1998) Pathways for protein folding: is a new view needed? *Curr. Opin. S1truct. Biol.* **8,** 68–79.

28. Krivov, S. and Karplus, M. (2004) Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. USA* **101,** 14766–14770.
29. Settanni, G., Rao, F., and Caflisch, A. (2005) Value analysis by molecular dynamics simulations of reversible folding. *Proc. Natl. Acad. Sci. USA* **102,** 628–633.
30. Best, R. and Hummer, G. (2005) Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA* **102,** 6732–6737.
31. Chandler, D. (1978) Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.* **68,** 2959–2970.
32. Northrup, S. H., Pear, M. R., Lee, C. Y., McCammon, J. A., and Karplus, M. (1982) Dynamical theory of activated processes in globular proteins. *Proc. Natl. Acad. Sci. USA* **79,** 4035–4039.
33. Fersht, A. R. and Sato, S. (2004) Value analysis and the nature of protein-folding transition states. *Proc. Natl. Acad. Sci. USA* **101,** 7976–7981.
34. Cavalli, A., Ferrara, P., and Caflisch, A. (2002) Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins* **47,** 305–314.
35. Sanchez, I. E. and Kiefhaber, T. (2003) Origin of unusual values in protein folding: evidence against specific nucleation sites. *J. Mol. Biol.* **334,** 1077–1085.
36. Fersht, A. R. (2004) Relationship of Leffler (Bronsted) values and protein folding values to positions of transition-state structures on reaction coordinates. *Proc. Natl. Acad. Sci. USA* **101,** 14338–14342.
37. Baldwin, R. L. and Rose, G. D. (1999) Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* **24,** 77–83.
38. Karplus, M. and Weaver, D. L. (1976) Protein folding dynamics. *Nature* **260,** 404–406.
39. Daggett, V., Li, A. J., Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1996) Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* **257,** 430–440.
40. Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1995) The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein-folding. *J. Mol. Biol.* **254,** 260–288.
41. Li, A. J. and Daggett, V. (1996) Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *J. Mol. Biol.* **257,** 412–429.
42. Settanni, G., Gsponer, J., and Caflisch, A. (2004) Formation of the folding nucleus of an SH3 domain investigated by loosely coupled molecular dynamics simulations. *Biophys. J.* **86,** 1691–1701.
43. Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. (2001) Three key residues form a critical contact network in a protein folding transition state. *Nature* **409,** 641–645.