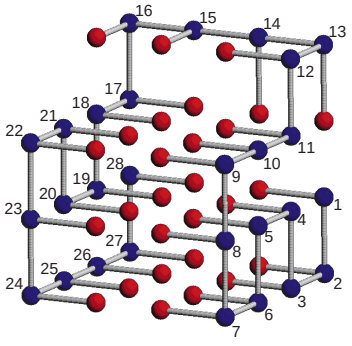
Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus

Lewyn Li, Leonid A. Mirny and Eugene I. Shakhnovich

A lattice model with side chains was used to investigate protein folding with computer simulations. In this model, we rigorously demonstrate the existence of a specific folding nucleus. This nucleus contains specific interactions not present in the native state that, when weakened, slow folding but do not change protein stability. Such a decoupling of folding kinetics from thermodynamics has been observed experimentally for real proteins. From our results, we conclude that specific non-native interactions in the transition state would give rise to ϕ-values that are negative or larger than unity. Furthermore, we demonstrate that residue Ile 34 in src SH3, which has been shown to be kinetically, but not thermodynamically, important, is universally conserved in proteins with the SH3 fold. This is a clear example of evolution optimizing the folding rate of a protein independent of its stability and function.

The amino acid sequence of a protein determines its specific three-dimensional structure[1]. It is important to understand the basis for this determination so that we can eventually design proteins that adopt desired conformations. In the continuing quest to understand how a protein folds, computer simulations have proven invaluable. In general, there are two ways to investigate protein folding in simulations: either all atoms and interactions are modeled realistically, or the protein is simplified. While an all-atom representation yields the most detail[2–4], current technology severely limits the duration of such a simulation, as well as the number of times it can be repeated[2–4]. In most cases, therefore, it is still unfeasible to follow the entire folding process using an all-atom representation of the protein and solvent.

Because of this, simplified models of proteins have been used extensively in computer simulations[5–12]. One of the most popular is the cubic lattice model[5,7], in which each amino acid along a protein is represented as a bead on a string, and the string is placed on a cubic lattice with each lattice site containing at most one bead. Results obtained using this model have greatly improved understanding of protein folding by facilitating the interpretation of data on real proteins. The most prominent example is the idea of the specific folding nucleus, which states that once a number of specific residues have come into contact, the protein has surmounted its free energy barrier and will rapidly fold into its native conformation[5]. This was first demonstrated in the cubic lattice model by Abkevich *et al.*[5], and has subsequently been used to interpret folding in studies of proteins such as CI2 (ref. 13), CheY[14], SH3 (refs 15–18) and FKBP12 (ref. 19). However, the cubic lattice model does have obvious shortcomings; chief among them is that amino acid side chains are not represented.



NSNRWRTAIAWVCHFIHWHWWAASHTKT

Fig. 1 Native state and wild type sequence of the protein model. The abundance of Trp in the sequence is an artifact of the parameters used in design, and should not be compared to real proteins. This is because there is no direct correspondence between real amino acids and the ones in lattice models[37].

Here we report a protein model that includes side chains and is based on the cubic lattice model. This variation of the cubic lattice model has been investigated, albeit in much less detail, by Bromberg and Dill in two dimensions, and by Klimov and Thirumalai[6] and Hart and Istrail[21] in three dimensions. Our study reveals the novel feature that, in addition to some particular native contacts, the transition state (TS) of a protein contains specific interactions not found in the native state. In addition, weakening these non-native interactions, instead of accelerating folding, actually slows folding down, without affecting the stability of the native state. This decoupling of folding kinetics from thermodynamics has been observed for many real proteins[13–18,22,23], and has often been attributed to non-native interactions[15–18,22]; our work strongly supports this interpretation. Finally, to assess possible evolutionary implications of our findings, we have performed evolutionary analyses on proteins containing the SH3 domain fold. The residue Ile 34 in src SH3, and its equivalent in other SH3 domains, seems to be universally conserved. Independent of our analysis, the same residue has been implicated in non-native interactions in the TS of α-spectrin and src SH3 (refs 15–18). Since non-native interactions in the TS affect the folding rate much more than stability (as shown by our computer simulations), we suggest that Ile 34 has been con-

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

Correspondence should be addressed to E.I.S. *email: eugene@belok.harvard.edu*

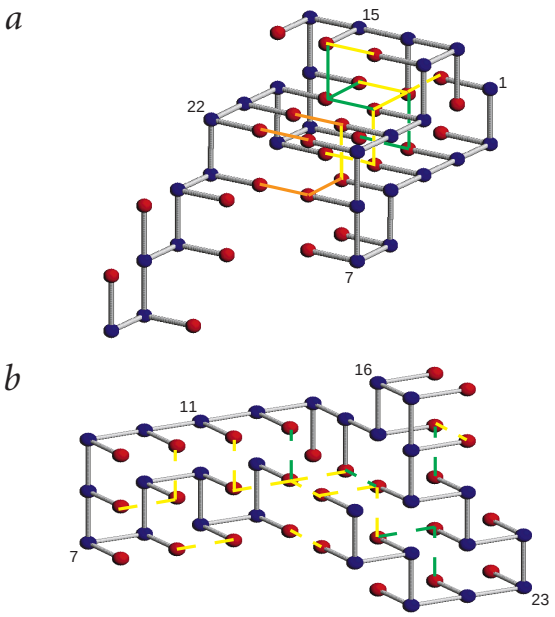


Fig. 2 Nucleus and control conformation of model protein. ***a***, Conformation with all nucleus contacts. Yellow and orange lines are the weak and strong native nucleus contacts, respectively, while green lines are non-native nucleus contacts. The native nucleus contacts are between monomers 1 and 14, 4 and 11, 4 and 19, 5 and 8, 5 and 10, 8 and 23, 9 and 22, 10 and 21, 11 and 14, 12 and 15, and 14 and 17. The non-native nucleus contacts are between monomers 3 and 14, 3 and 18, 11 and 20, 15 and 20, and 17 and 20. Monomers 1, 7, 15 and 22 have been labeled to aid identification. See text for definitions of weak and strong nucleus contacts. ***b***, Control conformation with 11 random non-local native contacts (yellow dashes) and 5 random non-local non-native contacts (green dashes). The native contacts are between monomers 1 and 4, 1 and 14, 1 and 28, 2 and 27, 3 and 6, 4 and 11, 5 and 8, 5 and 10, 15 and 18, 19 and 26, and 19 and 28. The non-native contacts are between monomers 1 and 12, 14 and 19, 15 and 20, 21 and 24, and 21 and 26. Monomers 7, 11, 16 and 23 have been labeled to aid identification.

served because of its kinetic, rather than functional or thermodynamic, importance.

This is, to the best of our knowledge, the first time such a decoupling between kinetics and thermodynamics has been reported for computer simulations and evolutionary analysis of protein folding.

Native state and folding nucleus of model protein

The native state and amino acid sequence of the protein model is shown in Fig. 1. It is a 28-mer with 51 contacts among the side chains. The model contains no interactions between backbone and backbone, or backbone and side chain, other than the requirement that they cannot occupy the same lattice site. Out of the 51 native contacts, 24 are between adjacent monomers (local contacts) and 27 are between non-adjacent monomers (non-local contacts). The native state energy is -7.24 (see Methods for units of energy) and, in over 100 independent runs, no conformation had an energy lower than -7.24. The sequence folded in a two-state manner at T = 0.07 (data not shown), which was the simulation temperature for all studies reported here, unless stated otherwise.

Of the non-local contacts, 11 native and 5 non-native contacts constitute the folding nucleus (see Methods for details). These are called 'nucleus contacts' and have been colored orange, yellow or green in Fig. 2*a* (see caption for their identities). A folding nucleus is defined as a conformation that, once achieved, leads to rapid and reproducible folding — that is, folded into the native state within 1×10^7 Monte Carlo steps (MCS) 75% of the time. Strictly speaking then, this is a 'post-critical' state rather than a transition state, but here we use the two terms interchangeably because they differ only by several kT[5]. On the other hand, the 24 native local contacts could not be clearly separated into nucleus and non-nucleus contacts (data not shown). This is probably due to the proximity of the two monomers in a local contact, which means that the contact can be easily satisfied, whether or not the rest of the protein is committed to folding.

To rigorously test the nucleus contacts, six different conformations that contained only the 16 nucleus contacts and 20–24 local native contacts were generated. In these conformations, the native state was typically achieved in less than 5×10^6 MCS — that is, less than 1% of the median first passage time (MFPT) of a random coil (Fig. 3). In contrast, the 6 control conformations (Fig. 2*b*), each having 11 random native non-local contacts, 5 random non-native non-local contacts, and 20–24 local native contacts, had a typical MFPT of 8.8×10^8 MCS — that is, of the same order of magnitude as a random coil (Fig. 3). This demonstrates conclusively that the 16 nucleus contacts are sufficient for rapid and reproducible folding and that the nucleus contacts are not random.

This is not to say, however, that all 16 nucleus contacts are necessary for rapid folding. In our database of putative nuclei, the 16 nucleus contacts all have a high probability (typically >0.5) of being present, but none of them was found in all the putative nuclei. Contacts other than the 16 selected were also detected in the database, albeit at a lower frequency (Fig. 4). It is plausible that some of them could occasionally substitute for one or two of the 16 nucleus contacts; therefore, the 16 nucleus contacts most likely represent the major component of the TS ensemble, around which there are small fluctuations. This picture of the TS is consistent with data from mutagenesis experiments[13–19,22–24].

The native nucleus contacts can be further classified as 'strong' and 'weak' according to their differential contact frequencies (DCFs; Fig. 4; see Methods). A highly positive DCF means that the contact appeared much more frequently in post-critical than pre-critical states, while a highly negative DCF means the opposite. A DCF near zero indicates that the contact has no preference for the post-critical or pre-critical state. A strong nucleus contact

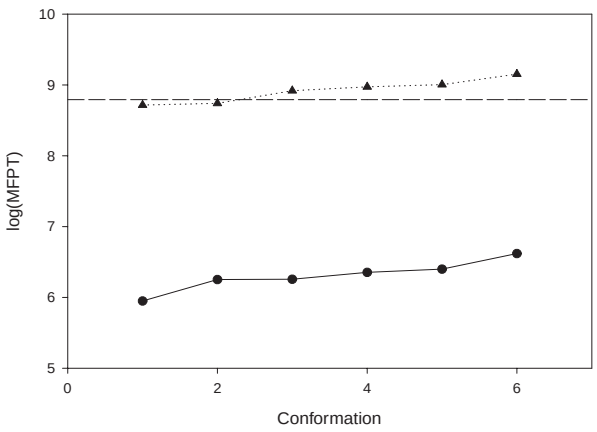


Fig. 3 The log(MFPT) of nucleus (circles), control (triangles) and random coil (dotted line). The line for random coil was drawn using data from Table 1. Each number on the abscissa represents one nucleus or control conformation, and has no other significance.

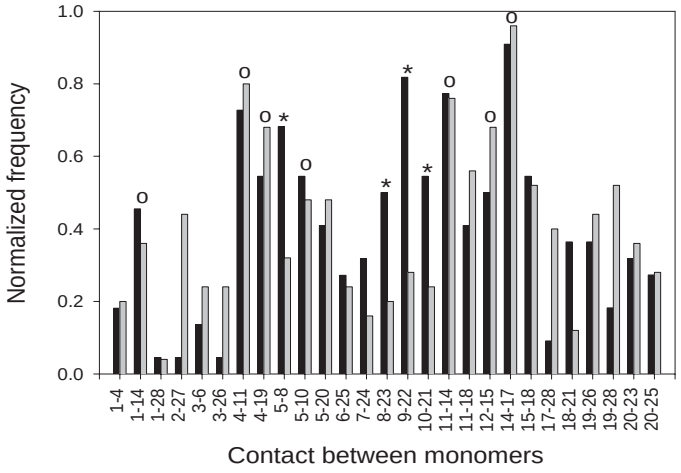


Fig. 4 Contact frequencies for non-local native contacts among post-critical (black bars) and pre-critical (gray bars) states. The label x-y on the abscissa refers to the contact between monomers x and y. Stars and circles represent strong and weak nucleus contacts, respectively. Note that contacts between monomers 2 and 27, 17 and 28, and 19 and 28 appeared much more frequently in pre-critical than post-critical states, so they may be 'trapping' contacts that discourage folding. See text and Methods for definitions of post- and pre-critical states, and of strong and weak nucleus contacts.

has a DCF ≥ 0.3, while a weak one has a DCF near zero (Fig. 4). The strong native nucleus contacts, colored orange in Fig. 2*a*, have an average energy of -0.13. The other native nucleus contacts are weak, with an average energy of -0.18. The strong native nucleus contacts are strong not because they are energetically more favorable, but instead may be so because of their positions; they form a cluster (Fig. 2*a*) and could be best placed to bring the two sides of the protein model together to form the native state. With DCFs near zero, the weak nucleus contacts are less specific for folding than the strong nucleus contacts. Their presence in the TS could be due to two reasons. First, the strong nucleus contacts might constrain the system so much that the weak nucleus contacts must form, as suggested by studies on α-spectrin SH3 (refs 15,16). Second, the weak nucleus contacts may confer the energetic stability necessary to make the TS accessible. We do not have sufficient data to repeat this analysis for the non-native nucleus contacts.

Non-native nucleus contacts

The five non-native nucleus contacts were mutated to make them all repulsive; four of them were attractive in the wild type sequence (Fig. 1; see Methods). For the control experiment, the five non-native nucleus contacts were the same as in the original sequence, but five other non-native contacts that were similar to the non-native nucleus contacts in energy and sequence separation were mutated. We did not alter the amino acids in the sequence, as in a real mutagenesis experiment, but merely made the nucleus or control contacts repulsive. This allowed us to study the influence of the contacts without drastically changing the folding pathway. Both sequences followed two-state kinetics at T = 0.07 (data not shown).

Kinetics and thermodynamics of folding

Previous computational studies on folding nuclei have focused on native contacts[5,25], and non-native contacts have rarely been explored[26]. *A priori*, one might expect that destabilizing non-native contacts would not significantly influence the folding rate because these contacts do not affect the energy of the native state. Our results for the control sequence conform to this expectation: the wild type sequence had a MFPT of $6.2 \pm 1.0 \times 10^8$ MCS, and the control sequence MFPT was almost the same at $6.3 \pm 1.0 \times 10^8$ MCS (Table 1).

Strikingly, the mutant sequence, in which the five nucleus non-native contacts were weakened, completely contradicted the above expectation and folded significantly more slowly than wild type, with a MFPT of $1.72 \pm 0.10 \times 10^9$ MCS (Table 1). This represents almost a three-fold deceleration, which was reproducible.

On the other hand, the thermodynamic stability of the native state was not noticeably affected by the various mutations. All three sequences showed normal sigmoidal denaturation when heated (data not shown) and had, to within experimental error, the same T_m (Table 1). However, the energy of the unfolded state ensemble did increase slightly in both the mutant and control. At T = 0.09, when the unfolded states dominated, the wild type had an average energy of -3.18 ± 0.01, while the mutant and the control had average energies of -3.12 ± 0.01 and -3.11 ± 0.01, respectively.

Comparison with experimental results

Currently, the most detailed method for probing the TS of a protein is site-directed mutagenesis and subsequent ϕ-value analysis[13–19,22–24]. A ϕ-value is defined as the ratio $(\Delta G_{U-\ddagger mutant} - \Delta G_{U-\ddagger wild\ type}) / (\Delta G_{U-F mutant} - \Delta G_{U-F wild\ type})$, where $\Delta G_{U-\ddagger}$ is the difference in free energy between the unfolded and TS, and ΔG_{U-F} is the difference in free energy between the unfolded and native state. If a protein folds in a two-state manner, a ϕ-value near unity means that a mutated residue participates fully in the TS, while a ϕ-value near zero indicates that it participates little[27]. A fractional ϕ-value can be due to partial participation in the TS or parallel pathways[28]. For most proteins ϕ-values fall between 0 and 1 (refs 13–19,22–24).

However, two categories of 'abnormal' mutations have also been observed[13–18,22,23]. The first category has a negative ϕ-value, which means that the mutation affects the TS in a manner opposite to its effect on the native state. Members of this category include L32I in CI2 (ref. 13), L33V of α-spectrin SH3 (refs 15,16) and W43I in src SH3 (refs 17,18). The second category consists of mutations that stabilize or destabilize both the TS and the native state, but affect the TS much more than the native state. This is usually manifested in a ϕ-value that is greater than unity, as in the case of V21T and N23G in the F14N mutant of CheY[14], I34A in src SH3 (refs 17,18), I23V in ADA2h[22] and Y25A in AcP[23]. The natural interpretation for these ϕ-values is that the residue takes part in non-native interactions[15–18,22]. In extreme cases, such as V21T/F14N in CheY[14] and I34A in src SH3 (refs 17,18), a mutation can reduce the refolding rate by an order of magnitude, while hardly affecting protein stability.

Table 1 Kinetic and thermodynamic parameters of the protein model

Name	MFPT (x 10^8 MCS)[1]	T_m[2]
Wild type	6.2 ± 1.0	0.065 ± 0.005
Mutant	17.2 ± 1.0	0.065 ± 0.005
Control	6.3 ± 1.0	0.065 ± 0.005

[1]Average ± uncertainty from two independent estimations of MFPT. Each estimation came from 55 simulations.
[2]Temperature at which <Q> = 0.55.

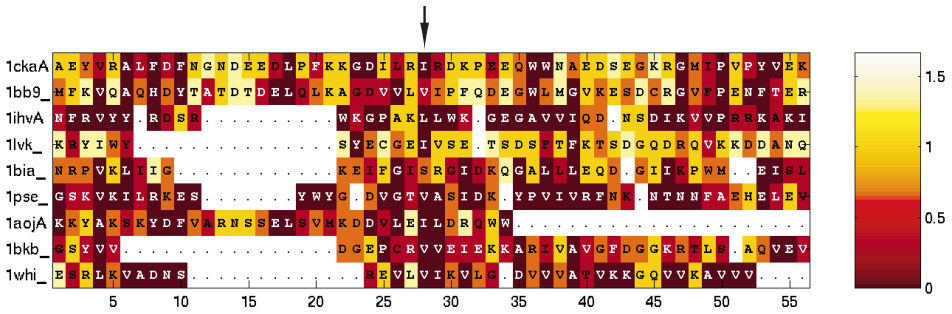


Fig. 5 Structural alignment of families containing the SH3 fold. Each line corresponds to one family of homologous proteins, and representative proteins for each family are indicated (see Methods for complete list). The color scale shows conservation within each family: brown, conserved; yellow, variable. For each family we include the PDB code and the sequence of the representative protein. Note that brown vertical stripes indicate universally conserved residues in most families. We used c-crk (1cka, chain A) as the representative of the SH3 domain, so our numbering differs from src SH3 (refs 17, 18) by six units for residues 1–42, and by seven units for residues 43–56; Ile 28 here is equivalent to Ile 34 in src SH3. The arrow points to Ile 34 in src SH3 and its equivalents in other proteins containing the SH3 fold.

The results reported here are in good qualitative agreement with the experimental evidence summarized above. The T_m of the mutant was identical to that of wild type to within experimental uncertainty, signifying little change in overall stability. Yet the same mutations lengthened the median folding time by over 150% (Table 1). These results therefore strongly support the interpretation that an 'abnormal' ϕ-value indicates non-native interactions in the TS. It was difficult to establish the sign of the 'ϕ-value' in the protein model since the change in stability was so small.

More importantly, these computational results illuminate many aspects of data from real proteins. The minor change in $\Delta G_{U\text{-}F}$ for mutants with 'abnormal' ϕ-values can be explained by considering how non-native nucleus contacts influence the native and unfolded states. Destabilizing non-native contacts, by definition, cannot change the energy of the native state. Such a destabilization can raise the energy of the unfolded state ensemble, as was found in our simulations. It can also reduce the entropy of the unfolded state ensemble by disfavoring conformations with the destabilized non-native contacts. Both of these will increase the stability of the native state. However, the effect is likely to be weak because the unfolded state, being an ensemble of disordered structures, has many different contacts, and destabilizing a few non-native contacts will not change the ensemble much. This would account for the constant T_m in our simulations, and is consistent with the small $\Delta\Delta G_{U\text{-}F}$ observed for many mutants with 'abnormal' ϕ-values.

These results also shed light on why the 'abnormal' mutations have a relatively strong influence on the refolding rate. Since the TS ensemble is much more structured than the unfolded state ensemble, it consists of fewer states[29]. Any destabilization of a structurally important contact, even if it is non-native, would make any TS structure containing the destabilized contacts energetically unfavorable. This would, in turn, drastically lower the entropy of the TS ensemble because the mutated system has far fewer energetically accessible pathways to the native state. As a result, the free energy of the TS ensemble will increase significantly, hence the slower refolding rate.

Of course, any mutation in a real protein changes the identity of the amino acid, so it could affect non-native and native contacts. This has not been taken into account here because, in our 'mutations', the specific non-native contacts were weakened but all native contacts were left untouched. Nevertheless, in a real protein, some residues may make important non-native contacts in the TS without participating strongly in the native state. The results and conclusions reported here probably apply more to this class of residues than to others.

Evolutionary optimization of folding in SH3 domains

The mutation I34A in src SH3 reduced the folding rate by almost an order of magnitude but destabilized the protein by only 0.32 kcal mol^{-1} (ref. 18). This gave a ϕ-value of 3.9 (ref. 18). Martinez *et al.* noted that Leu 33 in α-spectrin SH3, the structural equivalent to Ile 34 in src SH3, also had an 'abnormal' (negative) ϕ-value[15]. This agreement could have evolutionary relevance, especially as the homology between the two sequences is only 36%[15]. To explore this issue, we studied the sequence conservation in nine non-homologous proteins sharing the SH3 fold. A multiple sequence alignment for each protein was built first. Then the 'conservation of conservation' (S^{CoC}) and the conservation across the families (S^{across}) were determined (see Methods). A low S^{CoC} means that a position is conserved within each family, but the identity of the conserved residue may vary from family to family[30,31]. A low S^{across} indicates that most families have residues of the same type in that position[30,31].

Our analysis revealed that the amino acid at position Ile 34 in src SH3, and its analogs, is universally conserved, with low S^{across} and S^{CoC} (Figs 5, 6). Two observations suggest that this conservation is not due to protein function. First, Ile 34 is far from the peptide binding pocket of src SH3 (Fig. 6*a*). Second, Ile 34 and its equivalent (Leu 33 in α-spectrin SH3) are universally conserved across proteins of different functions (Figs 5, 6*b*). Nor are thermodynamics likely to be responsible for the conservation of Ile 34, as mutations of that residue had only a weak effect on protein stability[18]. This position, therefore, appears to be kinetically relevant, and its conservation points to possible evolutionary pressure on folding kinetics. The same pressure might have applied to Trp 43, which is also at a distance from the peptide binding pocket. The mutation W43I more than doubled the folding rate of src SH3, but destabilized it by only 0.77 kcal mol^{-1}, yielding a ϕ-value of -0.68 (ref. 18). Similar to Ile 34, Trp 43 is also conserved across families, albeit to a lesser extent (Figs 5, 6).

What is remarkable here is that the pressure seems to originate purely from kinetics; Ile 34 and Trp 43 may have been selectively conserved because of their ability to stabilize the TS through non-native interactions, and not because of their contribution to protein stability or function. This hypothesis could be tested by mutating positions equivalent to Ile 34 and Trp 43 of src SH3 in proteins that differ functionally from SH3 but have

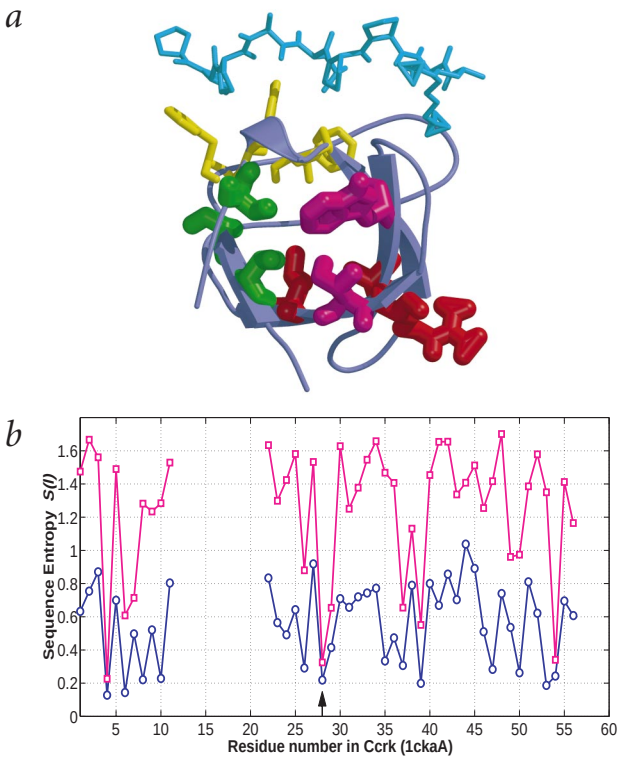


Fig. 6 Conserved residues in SH3. ***a***, Structure of c-crk SH3 with a proline-rich peptide, drawn using Molscript[38] and Raster3D[39]. Residues with high conservation of conservation (that is, low S^{CoC}) and high conservation across the families (that is, low S^{across}) are shown in thick wire-frame representation. Residues colored in magenta are those with abnormal ϕ-values in src SH3[17,18]: Ile 28 (Ile 34) and Trp 37 (Trp 43). Residues Leu 26 (Leu 32), Arg 29 (Val 35) and Ala 39 (Ala 45), all with high ϕ-values[17,18], are colored in red. Residues with low ϕ-values[17,18] are colored in green: Val 4 (Phe 10), Ala 6 (Ala 12) and Val 54 (Val 61). The notation Ile 28 (Ile 34) means Ile 28 in c-crk SH3 corresponds to Ile 34 in src SH3. Residues with low S^{CoC} but high S^{across} are shown in yellow thin wire-frame representation. Residues Phe 8 (Tyr 14), Phe 10 (Tyr 16), Pro 50 (Pro 57) and Tyr 53 (Tyr 60) all contribute to the peptide binding pocket. See Methods for details on how CoC and cross-family conservation were determined. ***b***, Conservation in proteins containing the SH3 fold. The magenta line indicates conservation across families, while the blue line indicates 'conservation of conservation' (CoC). The numbering is the same as in Fig. 5. The arrow points to Ile 34 in src SH3 and its equivalents in other proteins containing the SH3 fold. See Methods for details on how CoC and cross-family conservation were determined.

the same overall fold. 'Abnormal' ϕ-values for these mutations would support our hypothesis. Such an experiment has not been done to the best of our knowledge and may be of interest. Of course, this does not mean that the protein has been selected by evolution to be the fastest folder possible. Rather, we believe that folding must be sufficiently optimized for the protein to survive and function.

Other noteworthy features are apparent in Figs 5, 6. First, six residues exhibit low S^{CoC} and S^{across}. Of these, Leu 26 (Leu 32), Arg 29 (Val 35) and Ala 39 (Ala 45) participate in the SH3 folding nucleus (residues in parentheses are according to the numbering used by Baker and colleagues[17,18] for src SH3). The remaining three (Val 4 (Phe 10), Ala 6 (Ala 12) and Val 54 (Val 61)) belong to the hydrophobic core but not to the folding nucleus, as seen from their low ϕ-values[18]. Second, and in contrast, the nucleation residue Ile 49 (Ile 56) has a high S^{CoC} but a low S^{across}. Third, several polar amino acids in src SH3, such as Asp 41(Ser 47), have high ϕ-values but are not conserved (Fig. 6*b*); these residues may be specific to the folding nucleus of src SH3. Lastly, Phe 8 (Tyr 14), Phe 10 (Tyr 16), Pro 50 (Pro 57) and Tyr 53 (Tyr 60) have low S^{CoC} but are not conserved across families; these residues bind the proline-rich peptide (Fig. 6*a*) from the guanine nucleotide exchange factor C3G. Remarkably, their structural counterparts in another SH3-like barrel, the DNA-binding domain of HIV-1 integrase (PDB entry code 1IHV), are the conserved but basic residues Arg and Lys, which bind DNA.

Implications for protein folding, design and evolution

We believe that non-native nucleus contacts assist folding by providing energetically accessible pathways to the native state, pathways that cannot form through native contacts alone. Disfavoring such non-native contacts would slow folding down significantly. This has several implications for protein folding and design. First, in theoretical and experimental design of fast folding proteins, the non-native nucleus contacts, if they exist, should not be weakened. To establish the existence of such non-native contacts, we suggest that they should have ϕ-values that are negative or larger than unity. Second, if a folding nucleus contains no non-native nucleus contacts, it may be possible, by careful mutations, to introduce such contacts into the folding nucleus to accelerate folding. Third, since a non-native nucleus contact participates in the TS ensemble but not in the native state, its influence on protein stability might be weak. Mutating such contacts could therefore serve as a way to fine-tune folding kinetics without affecting thermodynamics. Finally, evolution may have conserved non-native interactions in the nucleus to optimize folding, as suggested by the analysis here.

Methods

General features of the model. The model has been described in detail elsewhere[6]. There were no interactions between side chain and backbone, or backbone and backbone, except chain connectivity and excluded volume. For the interactions between different side chains, data from Table 6 of Miyazawa and Jernigan[32] were used. In this study, we used the same energy unit as in ref. 32.

Monte Carlo kinetics. All simulations were done at T = 0.07 and began with a random conformation unless stated otherwise. At the beginning of each MCS, the energy of the system was evaluated and stored. One backbone atom was then selected at random and moved while preserving chain connectivity. The possible moves were tail flip (20% of the time), corner flip (20%) and crankshaft (60%). Once the backbone had been moved, one random site was selected for the side chain of the moved monomer. The Metropolis criterion[33] was applied to accept or reject the move. All moves violating excluded volume had infinite energy and were automatically rejected. In addition, one random side chain was chosen every 100 MCS, rotated once and accepted or rejected according to the Metropolis criterion. The Metropolis criterion states that if a move lowers or maintains the system energy, it is accepted. On the other hand, if a move increases the system energy, the Metropolis criterion accepts the move with probability = $\exp(-\Delta E/kT)$, where $\Delta E = E_{(\text{after move})} - E_{(\text{before move})}$.

Monitoring of proximity to the native state. To estimate how close to the native state our system was we used the quantity Q, defined as Q = N'/N where N' is the number of non-local native contacts at any MCS, and N is the number of non-local native contacts in the native state (27 in the case used here). A non-local contact is a contact between i and i + k residues, with k > 1. We excluded local (k = 1) native contacts in Q because they were often present even in the unfolded state ensemble (see text for discussion). We also made use of the χ_{bb} parameter defined by Klimov and Thirumalai[6] to monitor the backbone of our model protein.

Sequence design and mutations. The wild type sequence in Fig. 1 was designed using published procedures[34]. This procedure maximized the energy difference between the native state and the misfolds that competed with the native state for low energy[34]. It also minimized the energy dispersion among the native state contacts to ensure cooperative folding[34].

To destabilize the non-native nucleus contacts, a mutant sequence was created by making the contacts between monomers 3 and 14, 3 and 18, 11 and 20, 15 and 20, and 17 and 20 repulsive to the extent of +0.10. For the control sequence, the contacts between monomers 6 and 17, 6 and 21, 8 and 15, 9 and 16, and 14 and 21 were chosen because they were similar in energy and sequence separation to the non-native nucleus contacts. They were then made repulsive to the same extent of +0.10.

Determination of MFPT and T_m. For the MFPT, we ran 110 independent simulations, starting with random conformations and terminated when the backbone became fully native. We then separated the runs into two batches of 55 runs to test for reproducibility. FPT is the number of MCSs needed for the backbone of the model protein, starting from a random coil, to become fully native. The average MFPT and uncertainty are reported in Table 1. T_m was defined as the temperature at which <Q> = 0.55, which was the transition mid-point. <Q> at each temperature was obtained by averaging 10 independent runs. Each run started with the native state and lasted for 3×10^9 MCS.

Identification of the specific folding nucleus. First, we ran over 100 independent simulations and collected all the states with Q = 0.41 that were less than 1×10^7 MCS from the native state. We called these putative folding nuclei (PFN). We wanted to identify the minimal nucleus, so Q was kept as low as possible. But we also wanted a substantial number of PFNs to lend confidence to the statistical analysis. By trial and error, we found Q = 0.41 to be a good compromise as a selection cut-off. The PFNs may, or may not, contain the whole folding nucleus. This is because the nucleus could form early or late, and it is possible that for some runs only part of the nucleus is present at Q = 0.41 in a PFN, with the remainder forming slightly later. The only requirement for a folding nucleus is that once it is formed, the system rapidly and reproducibly descends to the native state.

We then launched 10–20 independent trajectories from each PFN. The conformations that folded rapidly and reproducibly (that is, in $<1 \times 10^7$ MCS for more than 75% of the runs) were designated 'post-critical' states, while the ones folding rapidly in no more than 10% of the runs were designated 'pre-critical'. The frequencies of non-local native contacts in both groups were then computed. The post-critical and pre-critical states showed qualitatively different distributions of contact frequencies (data not shown). The contacts with probability >0.5 in the post-critical states were chosen as the native nucleus contacts. The only exception was the contact between 15 and 18. This was because (as contacts between 11 and 14, and between 12 and 15 were already in the nucleus) topological constraint made it impossible to include the contact between 15 and 18 without forming an extra native contact between 11 and 18. As the contact between 11 and 18 was not among the contacts with probability >0.5, including the contact between 15 and 18 would artificially introduce the contact between 11 and 18 into the nucleus. The contact between 15 and 18 was therefore replaced with a contact between 1 and 14, which was the next most frequently encountered native contact among the post-critical states. Similar analyses were performed for the local native contacts and the non-native contacts, which identified the five non-native nucleus contacts.

The differential frequency for each native contact was determined by subtracting the frequency of the contact appearing in the pre-critical states from the frequency in the post-critical states, using the data in Fig. 4.

Analysis of conservation in SH3 folds. This was carried out as described[30,31]. Nine representative but non-homologous proteins with SH3-like folds were used: c-crk (PDB entry code 1cka, chain A), amphiphysin (PDB entry code 1bb9), HIV-1 integrase (PDB entry code 1ihv, chain A), myosin S1 fragment (PDB entry code 1lvk), Bira bifunctional protein (PDB entry code 1bia), photosystem I accessory protein E (PDB entry code 1pse), eps8 fragment (PDB entry code 1aoj, chain A), transcription initiation factor (PDB entry code 1bkb), and ribosomal protein L14 (PDB entry code 1whi). First, multiple sequence alignments of each protein with its homologs were obtained to form a family[35]. The conservation was then computed within each family as $s_i^m = -\sum_{b=1}^{x} f_i^m(b) \log f_i^m(b)$ where $f_i^m(b)$ is the frequency of residue type b at position i of the family m. The nine representative proteins were then structurally aligned with each other[36]. The conservation of conservation (S^{CoC}) at position i in the multiple structural alignment was then calculated as $s_i^{CoC} = 1/M \sum_{m=1}^{M} s_i^m$, with M = 9. Conservation across all the families was also measured as $S_i^{across} = -\sum_{b=1}^{x} F_i(b) \log F_i(b)$ where $F_i(b) = 1/M \sum_{m=1}^{M} f_i^m(b)$ is the average frequency of residue type b at position i among all the families. This averaging is important to make all families contribute equally, regardless of the number of homologs in each family. The residues were grouped into six classes according to their physical properties, as in (refs 30,31),

Acknowledgments

We thank the NSERC of Canada and NIH for financial support. L.L. would like to thank H. Angerman, G. Berriz, S. Choe, N.V. Dokholyan, L. Gutman, A. Ishchenko, E. Kussell, M. Morrissey and J. Shimada for computer assistance and stimulating discussions. He is particularly grateful to O. Clement for encouragement at a crucial point in this research.

Received 20 December, 1999; accepted 3 March, 2000.

1. Creighton, T.E. *Proteins: structures and molecular properties, 2nd edition* (W.H. Freeman and Company, New York; 1993).
2. Li, A. & Daggett, V. Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate. *J. Mol. Biol*. **275**, 677–694 (1998).
3. Duan, Y. & Kollman, P.A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744 (1998).
4. Lazaridis, T. & Karplus, M. 'New view' of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278**, 1928–1931 (1997).
5. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. Specific nucleus as the transition state for protein folding: evidence from lattice model. *Biochemistry* **33**, 10026–10036 (1994).
6. Klimov, D.K. & Thirumalai, D. Cooperativity in protein folding, from lattice models with sidechains to real proteins. *Folding Design* **3**, 127–139 (1998).
7. Socci, N.D., Onuchic, J.N. & Wolynes, P.G. Protein folding mechanisms and the multidimensional folding funnel. *Proteins* **32**, 136–158 (1998).
8. Hao, M-H & Scheraga, H.A. Molecular mechanisms for cooperative folding of proteins. *J. Mol. Biol*. **277**, 973–983 (1998).
9. Kolinski, A. & Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* **18**, 338–352 (1994).
10. Berriz, G.F., Gutin, A.M. & Shakhnovich, E.I. Cooperativity and stability in a Langevin model of proteinlike folding. *J. Chem. Phys.* **106**, 9276–9285 (1997).
11. Guo, Z. & Brooks III, C.L. Thermodynamics of protein folding: a statistical mechanical study of a small all-β protein. *Biopolymers* **42**, 745–757 (1997).
12. Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E. & Shakhnovich, E.I. Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol*. **296**, 1183–1188 (2000).
13. Itzhaki, L.S., Otzen, D.E. & Fersht, A.R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol*. **254**, 260–288 (1995).
14. López-Hernández, E. & Serrano, L. Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Folding Design* **1**, 43–55 (1996).
15. Martinez, J.C., Pisabarro, M.T. & Serrano, L. Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721–729 (1998).
16. Martinez, J.C. & Serrano, L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct.Biol.* **6**, 1010–1015 (1999).
17. Grantcharova, V.P., Riddle, D.S., Santiago, J.V. & Baker, D. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714–720 (1998).
18. Riddle, D.S. *et al*. Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016–1024 (1999).
19. Fulton, K.F., Main, E.R.G., Daggett, V. & Jackson, S.E. Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol*. **291**, 445–461 (1999).
20. Bromberg, S. & Dill, K.A. Side-chain entropy and packing in proteins. *Protein Sci.* **3**, 997–1009 (1994).
21. Hart, W.E. & Istrail, S. Lattice and off-lattice side chain models of protein folding: linear time structure prediction better the 86% of optimal. *J. Comp. Biol.* **4**, 241–259 (1997).
22. Villegas, V., Martínez, J.C., Avilés, F.X. & Serrano, L. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol*. **283**, 1027–1036 (1998).
23. Dobson, C.M. *et al*. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005–1009 (1999).
24. Daggett, V., Li, A. & Fersht, A.R. Combined molecular dynamics and Φ-value analysis of structure-reactivity relationships in the transition state and unfolding pathway of barnase: structural basis of Hammond and anti-Hammond effects. *J. Am. Chem. Soc.* **120**, 12740–12754 (1998).
25. Klimov, D.K. & Thirumalai, D. Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol*. **282**, 471–492 (1998).
26. Gutin, A.M., Abkevich, V.I. & Shakhnovich, E.I. A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Folding Design* **3**, 183–194 (1998).
27. Matouschek, A., Kellis, J.T., Serrano L. & Fersht, A.R. Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340**, 122–126 (1989).
28. Fersht, A.R., Itzhaki, L.S., ElMasry, N.F., Matthews, J.M. & Otzen, D.E. Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *Proc. Natl Acad. Sci. USA* **91**, 10426–10429 (1994).
29. _ali, A., Shakhnovich, E. & Karplus, M. How does a protein fold? *Nature* **369**, 248–251 (1994).
30. Mirny, L.A., Abkevich, V.I. & Shakhnovich, E.I. How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA* **95**, 4976–4981 (1998).
31. Mirny, L.A. & Shakhnovich, E.I. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol*. **291**, 177–196 (1999).
32. Miyazawa, S. & Jernigan, R. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).
33. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
34. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. Improved design of stable and fast-folding model proteins. *Folding Design* **1**, 221–230 (1996).
35. Dodge, C., Schneider, R. & Sander, C. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.* **26**, 313–315 (1998).
36. Holm, L. & Sander, C. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* **24**, 206–209 (1996).
37. Shakhnovich, E.I. Folding nucleus: specific or multiple? Insights from lattice models and experiments. *Folding Design* **3**, R108–R111 (1998).
38. Per, J.K. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. App. Crystallogr.* **24**, 946–950 (1991).
39. Merrit, E.A. & Bacon, D.J. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).