Side chain burial and hydrophobic core packing in protein folding transition states

PATRICK J. FARBER AND ANTHONY MITTERMAIER

Department of Chemistry, McGill University, Montreal H3A 2K6, Canada

(RECEIVED July 3, 2007; FINAL REVISION July 3, 2007; ACCEPTED December 28, 2007)

Abstract

A critical step in the folding pathway of globular proteins is the formation of a tightly packed hydrophobic core. Several mutational studies have addressed the question of whether tight packing interactions are present during the rate-limiting step of folding. In some of these investigations, substituted side chains have been assumed to form native-like interactions in the transition state when the folding rates of mutant proteins correlate with their native-state stabilities. Alternatively, it has been argued that side chains participate in nonspecific hydrophobic collapse when the folding rates of mutant proteins correlate with side-chain hydrophobicity. In a reanalysis of published data, we have found that folding rates often correlate similarly well, or poorly, with both native-state stability and side-chain hydrophobicity, and it is therefore not possible to select an appropriate transition state model based on these one-parameter correlations. We show that this ambiguity can be resolved using a two-parameter model in which side chain burial and the formation of all other native-like interactions can occur asynchronously. Notably, the model agrees well with experimental data, even for positions where the one-parameter correlations are poor. We find that many side chains experience a previously unrecognized type of transition state environment in which specific, native-like interactions are formed, but hydrophobic burial dominates. Implications of these results to the design and analysis of protein folding studies are discussed.

Keywords: protein folding; phi value analysis; hydrophobic core; hydrophobicity; side chain packing; mutagenesis

Supplemental material: see www.proteinscience.org

Proteins that fold with two-state kinetics are widely studied as model systems for the folding process. Although these molecules adopt their native conformations without the detectable accumulation of partly structured intermediates, information on rate-determining steps during the reaction can be obtained by comparing the effects of mutations on folding kinetics and native-state stability.

In a typical analysis, a parameter Φ is calculated for each mutation (Matouschek et al. 1989) according to

$$\Phi = \frac{-RT(\ln k_F^{mut} - \ln k_F^{wt})}{G_N}, \quad (1)$$

where k_F^{wt} and k_F^{mut} are the folding rates of the wild-type and mutant proteins and G_N is the change in the stability of the native state caused by the mutation,

$$G_N = \Delta G_{U \rightarrow N}^{mut} - \Delta G_{U \rightarrow N}^{wt}. \quad (2)$$

If it is assumed that protein folding follows Arrhenius kinetics with a mutation-independent prefactor, then the numerator in Equation 1 is equal to the change in the stability of the transition state,

Reprint requests to: Anthony Mittermaier, Department of Chemistry, McGill University, 801 Sherbrooke Street West, Room 322, Montreal H3A 2K6, Canada; e-mail: anthony.mittermaier@mcgill.ca; fax: (514) 398-3797.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.073105408>.

$$G_T \equiv \Delta G_{U \rightarrow \ddagger}^{\text{mut}} - \Delta G_{U \rightarrow \ddagger}^{\text{wt}} = -RT(\ln k_F^{\text{mut}} - \ln k_F^{\text{wt}}). \quad (3)$$

Mutations of side chains that are unstructured in the transition state produce Φ values of 0, while mutations of side chains that are natively structured in the transition state produce Φ values of 1. Intermediate values of Φ are obtained when interactions are only partially formed in the transition state or when folding proceeds through parallel transition states of which only a subset contain native-like interactions (Matouschek et al. 1989). Discrimination between single and parallel pathways can be achieved with Brønsted analyses, in which changes in transition-state stability, G_T , are plotted as a function of changes in native stability, G_N , for multiple mutations (Fersht et al. 1994). These plots can also provide evidence for the existence of native-like interactions in the transition state: When G_T depends linearly on G_N ,

$$G_T = \Phi G_N, \quad (4)$$

Φ is taken as the fraction of native contacts formed in transition state of a unique folding pathway (Fersht et al. 1994).

Φ value analysis has been used to investigate whether the tightly packed hydrophobic core characteristic of native globular proteins is formed in the transition state or at some later point in the folding process. In some cases, Brønsted analyses suggest that native-like hydrophobic core packing is present in the transition state. For example, in the chymotrypsin inhibitor CI2, folding data for multiple mutations at three buried positions in the “minicore” region of the protein produce a linear Brønsted plot with a slope of 0.3 (Itzhaki et al. 1995). This suggests that within the transition state, all minicore side chains experience interactions similar to those present in the folded state, but weakened by two-thirds, on average. Recently, a different relationship has been reported for the folding rates of hydrophobic core mutants. G_T values obtained for multiple substitutions at individual positions in the N-terminal domain of ribosomal protein L9 (NTL9) and the SH3 domain from the Fyn tyrosine kinase (FYN) correlate with changes in side-chain hydrophobicity, such that

$$G_T = \eta G_H, \quad (5)$$

where G_H is the change in side-chain hydrophobicity as calculated from experimental or theoretical water/octanol transfer free energies, $G_H = \Delta G_{\text{transfer}}^{\text{mut}} - \Delta G_{\text{transfer}}^{\text{wt}}$, and η is a position-specific constant of proportionality (Northey et al. 2002a; Anil et al. 2005). The implication is that side chains at these positions do not form specific native-like interactions in the transition state and instead influence transition-state stability purely through hydrophobic burial. Larger, more hydrophobic side chains

provide greater transition-state stability and faster folding rates without the specific shape requirements that characterize packing in the folded state (Anil et al. 2005).

If folding data agree with Equation 4 but not with Equation 5, or vice versa, it is possible to conclude that a position is involved in a subset of native-like interactions or, conversely, in nonspecific hydrophobic collapse in the transition state. However, in most cases, G_T values correlate with both G_N and G_H , so it is not possible to unambiguously assign a core position to either of the two models. This is perhaps not surprising, since core mutations that reduce side-chain hydrophobicity also tend to destabilize the native state. Another issue in interpreting the folding rates of hydrophobic core mutants is that correlations between G_T and both G_H and G_N can be quite poor. One hypothesis is that disruptive mutations have perturbed the structure of the transition state such that the folding pathways of the wild-type and mutant proteins are significantly different (Northey et al. 2002a). As described below, we offer an alternative explanation.

We have investigated a model in which shape-specific, native-like interactions are present at the rate-limiting step of folding, but do not necessarily form simultaneously with side chain burial. Taking an approach similar to one used by Dill and coworkers to investigate α -helix formation in transition states (Weikl and Dill 2007), we fit data from multiple mutations at individual hydrophobic core positions in CI2, NTL9, and FYN to an equation that accounts for both native-like interactions and additional hydrophobic collapse,

$$G_T = \chi_N G_N + \chi_H G_H. \quad (6)$$

The first term corresponds to both hydrophobic burial and specific packing interactions that are shared by the native and transition states. The second term accounts for additional, i.e., nonnative, contributions of hydrophobic burial to the stability of the transition state. The changes in the stabilities of the native and transition states for each mutant, G_N and G_T , were taken from published experimental data, and the changes in side-chain hydrophobicity, G_H , were calculated based on experimental or theoretical octanol/water transfer free energies (Hansch and Leo 1979; Fauchere and Pliska 1983). Using Equation 6, the separate contributions of native-like interactions and additional hydrophobic collapse to the stability of the transition state can be evaluated individually in the form of χ_N and χ_H parameters. In this way it is possible to unambiguously determine whether native-like core interactions are present at the rate-limiting step of folding. Furthermore, we show that the weakness of correlations between G_T and G_N observed for some positions is not due to significant disruption of the folding pathway. The experimental data are well fit by Equation 6, suggesting

that all mutants at these positions fold via similar pathways with similar rate-limiting steps at which side chain burial and the formation of native-like interactions play separate but important roles.

Theory

The parameters χ_N and χ_H provide specific information regarding transition state interactions for the side chain at a given position in the primary amino acid sequence. In cases where the position is fully solvent exposed in the unfolded state and fully buried in the native state, the sum ($\chi_N + \chi_H$) corresponds approximately to the extent of side chain burial in the transition state. Even when the position is not fully buried and exposed in the folded and unfolded states, respectively, nonzero values of χ_H arise only when side chain burial is asynchronous with the formation of native-like packing interactions. This follows from a consideration of the change in native-state stability resulting from a hydrophobic core mutation, G_N , which may be separated into two components,

$$G_N = G_{reorg} + G_{burial}. \quad (7)$$

G_{burial} refers to the contribution of changes in side-chain hydrophobicity to the stabilization or destabilization of the native state and G_{reorg} contains all other contributions to G_N , including changes in side chain packing, geometric strain, hydrogen and electrostatic bonding, and conformational entropy in both the native and unfolded states (Eriksson et al. 1992; Richards and Lim 1994). When a position is fully exposed in the unfolded state and fully buried in the folded state, $G_{burial} = G_H$, and

$$G_T = \chi_N G_{reorg} + (\chi_N + \chi_H) G_H. \quad (8)$$

The parameter χ_N can be interpreted qualitatively as the fraction of the specific, native interactions contributing to G_{reorg} that are formed in the transition state, including both local and long-range effects, while $(\chi_H + \chi_N)$ can be interpreted as the fractional burial of the side chain in the transition state.

Even in cases where a position is partially solvent exposed in the folded state or partially buried in the unfolded state, the parameter χ_H still provides clear information about the structure of the transition state. The fraction of the change in native-state stability that is due to changes in hydrophobic desolvation is given by

$$f_N^{hyd} = \frac{G_{burial}}{G_N}, \quad (9)$$

where G_{burial} may be less than or equal to G_H . Combining Equations 6 and 7, the fraction of G_T due to hydrophobic burial is given by

$$f_T^{hyd} = \frac{\chi_N G_{burial} + \chi_H G_H}{\chi_N G_N + \chi_H G_H}. \quad (10)$$

When $\chi_H > 0$, nonspecific hydrophobic interactions play a proportionally larger role in the stability of the transition state than they do in the native state and $f_T^{hyd} > f_N^{hyd}$, regardless of the solvent exposure of the position in the folded and unfolded states.

Results

We have analyzed data for three proteins that fold with two-state kinetics and where at least one hydrophobic core position has been substituted with four or more different amino acids: the chymotrypsin inhibitor 2 (CI2) (one position) (Itzhaki et al. 1995), the N-terminal domain of ribosomal protein L9 (NTL9) (four positions) (Anil et al. 2005), and the SH3 domain from the Fyn tyrosine kinase (FYN) (six positions) (Northey et al. 2002a; de los Rios et al. 2005). The one position in CI2 examined (Ile30) does not participate in the main hydrophobic core; however, it is largely buried in the protein interior and contacts several core residues (Itzhaki et al. 1995); therefore we believe that data for this position are suitable for the two-parameter analysis. For each position, we fit the parameters χ_N and χ_H to the set of changes in native-state stability, G_N , transition-state stability, G_T , and amino acid water/octanol transfer free energy, G_H , using Equation 6 and two-parameter linear regression through the origin. Values of χ_N , χ_H , and $(\chi_N + \chi_H)$ are listed in Table 1, together with 90% confidence intervals computed using a Monte Carlo approach (Efron and Tibshirani 1986). When the 90% confidence interval does not include zero, the value of χ_N , χ_H , or $(\chi_N + \chi_H)$ is considered to be statistically significant.

The motivation for this analysis was to extract information on side chain interactions at the rate-limiting step of folding, particularly in cases where it has not been possible to firmly establish whether specific, native-like interactions are present in the transition state. This is true for most positions used in this study. Pearson correlation coefficients between G_T and G_N , r_{TN} , and between G_T and G_H , r_{TH} , listed in Table 1, are both much larger than zero in most cases, and it is not possible to unequivocally select either Equation 4 or 5 as an appropriate model. For example, mutations of Ile28 in FYN produce r_{TN} and r_{TH} values of 0.89 and 0.83, respectively, and plots of G_T versus G_N and G_H , shown in Figure 1A,B are qualitatively similar. The G_T values correlate to some extent with both G_N and G_H , although deviations from the linear fits are much greater than the uncertainties in G_T . Significantly better agreement is obtained when χ_N and χ_H values are extracted from linear regression using Equation 6 and used to back-calculate G_T , as shown in Figure 1C.

Table 1. Two-parameter fits of folding kinetics

Protein	Position	n^a	r_{TN}^b	r_{TH}^c	χ_N^d	χ_H^e	$\chi_N + \chi_H^f$	Burial ^g
CI2	I30	4	0.97	0.94	0.20 (0.12,0.27)	0.13 (0.06,0.20)	0.33 (0.30,0.36)	0.78
NTL9	V3	5	0.01	0.95	−0.02 (−0.12,0.07)	0.80 (0.56,1.04)	0.78 (0.53,1.03)	1.00
	I4	4	0.97	0.95	0.19 (0.04,0.34)	0.24 (0.10,0.39)	0.43 (0.35,0.51)	0.98
	L6	5	0.99	0.97	0.16 (−0.07,0.38)	0.14 (−0.17,0.44)	0.29 (0.19,0.40)	0.78
	V21	4	0.86	0.95	0.03 (−0.05,0.12)	0.39 (0.24,0.55)	0.43 (0.29,0.57)	0.97
FYN	F4	4	0.70	0.80	0.12 (0.09,0.16)	0.09 (0.05,0.12)	0.21 (0.18,0.23)	0.91
	F20	4	0.78	0.89	−0.13 (−0.30,0.05)	0.35 (0.11,0.59)	0.23 (0.14,0.31)	0.99
	I28	5	0.89	0.83	0.38 (0.31,0.46)	0.39 (0.28,0.51)	0.78 (0.71,0.85)	0.92
	A39	5	0.76	0.58	0.61 (0.41,0.82)	0.58 (0.29,0.87)	1.19 (0.76,1.63)	1.00
	I50	4	0.39	0.90	0.05 (−0.06,0.16)	0.46 (0.23,0.69)	0.51 (0.33,0.70)	1.00
	V55	8	0.85	0.55	0.07 (−0.02,0.16)	0.00 (−0.15,0.16)	0.07 (−0.03,0.17)	0.96

^aNumber of mutants analyzed.^bPearson correlation coefficient for G_T and G_N .^cPearson correlation coefficient for G_T and G_H .^dDependence of changes in transition-state free energy on changes in native-state stability from fits of Equation 6. The 90% confidence interval is given in parentheses.^eDependence of changes in transition-state free energy on changes in side-chain hydrophobicity from fits of Equation 6. The 90% confidence interval is given in parentheses.^fFractional burial in the transition state for a position that is 0% buried in the unfolded state and 100% buried in the native state. The 90% confidence interval is given in parentheses.^gFractional burial of wild-type amino acid in the native state calculated using the MOLMOL software package (Koradi et al. 1996) and structures for CI2 (3CI2) (Ludvigsen et al. 1991), NTL9 (1DIV) (Hoffman et al. 1994), and FYN (1SHF) (Noble et al. 1993).

Mutations of FYN Ala39 produce much lower values of the correlation coefficients r_{TN} and r_{TH} (0.76 and 0.58), and Φ values range from −0.06 for A39F to 0.86 for A39G. Plots of G_T versus G_N , shown in Figure 1D, and G_T versus G_H , shown in Figure 1E, are not linear. Even in this case, fits using Equation 6 are able to closely reproduce experimental G_T values, as shown in Figure 1F. Notably, most of the Ala39 mutations were excluded from the original analysis, due to concerns that such disruptive, volume-increasing mutations could perturb the folding pathway (Northey et al. 2002a). The good agreement obtained using the two-parameter model suggests that all Ala39 mutants fold with similar mechanisms, passing through similar transition states. Overall, Equation 6 provides significant improvements in fit over Equations 4 and 5 for mutational data from CI2 Ile30, NTL9 Ile4, and FYN Phe4, Ile28, and Ala39.

When the analysis is extended to the 11 positions considered in this study, a high level of agreement is obtained; considering all 59 mutations, the root-mean-squared-deviation (RMSD) between back-calculated and experimental G_T values is 0.16 kcal/mol. In addition, we performed leave-one-out cross-validation (LOOCV) on a per-position basis to test the predictive ability of χ_N and χ_H . Data for each mutation was systematically removed to produce a validation data set, and two-parameter fitting was applied to the remaining $n - 1$ data points. Values of χ_N and χ_H thus obtained were used to predict the validation data set. The experimental and predicted validation data sets were in good agreement, with an overall correlation coefficient of 0.83 and an RMSD of 0.33 kcal/

mol. This confirms that Equation 6 is applicable to a wide array of mutations at different hydrophobic core positions in a variety of proteins. More importantly, it shows that the environments of hydrophobic core side chains in the transition state can be well described by just two parameters: one that depends only on the degree of hydrophobic burial and another that depends on the formation of all other native-like interactions.

Notably, values of χ_H for all positions considered, with the exception of NTL9 Leu6 and FYN Val55, are

Figure 1. Experimentally determined changes in transition-state free energies, G_T^{exp} , plotted as a function of changes in native-state stability, G_N (A,D), side-chain hydrophobicity, G_H (B,E), and back-calculated values computed using Equation 11, G_T^{calc} (C,F) for mutations of Ile28 (A,B,C) and Ala39 (D,E,F) in the Fyn SH3 domain. Equations for the lines in A, B, D, and E were obtained by one-parameter linear regression through the origin. Lines in C and F have a slope of 1 and a y-intercept of 0. Error bars corresponding to experimental uncertainties are smaller than the symbols shown.

significantly larger than zero. This implies that for essentially all positions considered here, the process of nonspecific hydrophobic collapse is more complete in the transition state than is the formation of native interactions. This result agrees with the nucleation–condensation model of protein folding in which the transition state is relatively compact even as native interactions are still in the process of being consolidated (Fersht 1999). In the cases where χ_N is not significantly greater than zero, data may be adequately fit by Equation 5, and contributions to transition-state stability are due primarily to nonspecific hydrophobic collapse. In contrast, when both χ_H and χ_N are significantly greater than zero, as is seen for five positions, the two-parameter model predicts that side chains participate in both specific native-like interactions and additional nonspecific hydrophobic collapse. Notably, this analysis shows that the interactions of a single side chain in the transition state often comprise a mixture of specific packing and nonspecific burial. Such mixed-mode interactions at the individual side chain level have not been previously considered in the interpretation of protein folding data.

The two-parameter analysis also provides an estimate of total side chain burial, via the sum ($\chi_H + \chi_N$). Three positions, NTL9 Val3 and FYN Ile28, Ala39 are highly buried with $(\chi_H + \chi_N) = 0.78, 0.78$, and 1.19 , respectively. The burial of NTL9 Val3 is primarily due to nonspecific collapse, since χ_N is essentially zero for this position, while for FYN Ile28 and Ala39, $\chi_N = 0.38$ and 0.61 , implying that significant native-like interactions are made in the transition state. Note that the value of $(\chi_H + \chi_N) = 1.19$ obtained for A39 does not differ significantly from the maximum expected value of $(\chi_H + \chi_N) = 1.0$, since 1.0 lies within the 90% confidence region calculated for this parameter and reported in Table 1.

There are two hydrophobic core positions for which neither χ_N nor χ_H is statistically significant. In the case of FYN Val55, a detailed analysis has shown that hydrophobic and hydrophilic substitutions produce quite different effects on the folding rate through a mechanism that is not well understood (de los Rios et al. 2005). It is therefore expected that G_T values would not correlate strongly with either G_N or G_H . For NTL9 Leu6, the sum $(\chi_H + \chi_N)$ is significantly larger than zero, implying that this position is partially buried, although neither χ_N nor χ_H is statistically significant on its own. This anomalous behavior is largely due to multicollinearity of native-state stability and side-chain hydrophobicity for this position. The correlation coefficient between G_N and G_H is 0.95 and the regression line passes nearly through the origin. Data for Leu6 agree with Equations 4, 5, and 6 to such an extent that it is not possible to discriminate specific from nonspecific interactions in this case, although it is clear from $(\chi_H + \chi_N)$ that contacts are formed by Leu6 in the transition state (see supplemental material).

Discussion

We have analyzed protein folding data using a two-parameter model that relates changes in folding rates to both changes in native-state stability and changes in side-chain hydrophobicity. This method provides information on the structure of the transition state that is not available from single-factor analyses of protein folding data. Based on correlations between folding rates and side-chain hydrophobicity, it had been proposed that side chains at a number of hydrophobic core positions in NTL9 and FYN contribute to the stability of the transition state purely via the hydrophobic effect in a non-shape-specific manner (Northey et al. 2002a; Anil et al. 2005). However, for several of these positions (NTL9 Ile4, Leu6; FYN Phe4, Phe20) non-negligible correlations also exist between folding rates and native-state stability, which suggests that some native-like interactions could be present in the transition state as well. Similarly, several mutations at positions believed to be partially natively structured in the transition state (Itzhaki et al. 1995; Northey et al. 2002a) produce folding rates that correlate with side-chain hydrophobicity (CI2 Ile30; FYN Ile28, Ala39, Ile50), raising the possibility that nonspecific hydrophobic collapse could play a dominant role. The two-parameter model resolves these ambiguities by effectively factoring out the effects of changes in side-chain hydrophobicity, G_H , and testing whether the dependence of G_T on G_N remains. In some cases, there is no significant correlation between G_T and G_N , once the dependence of G_T on G_H is taken into account. This provides a much stronger argument for the absence of shape-specific, native-like contacts in the transition state than simply a correlation between G_T and G_H . In other cases, G_T values show a significant dependence on G_N ($\chi_N > 0$), even after factoring out the influence of G_H , which clearly demonstrates that specific, native-like interactions are present in the transition state.

These results shed additional light on the folding mechanisms of the proteins studied here. For example, the good agreement between G_T and G_H values for NTL9 suggested that the transition state of this molecule is stabilized entirely by nonspecific hydrophobic collapse, with all other native-like interactions formed later in the folding process (Anil et al. 2005). The two-parameter fits employed here agree with this conclusion to some extent. Val3 and Val21 have significant χ_H values, indicative of nonspecific hydrophobic collapse, and nonsignificant values of χ_N , implying that side chains at these positions lack specific contacts in the transition state. However, χ_N is significantly greater than zero for Ile4, which suggests that it participates in specific, native-like contacts at the rate-limiting step of folding. These interactions could potentially involve Leu6. This position is immediately

adjacent in the native structure, as shown in Figure 2A, and may or may not be involved in specific interactions, as discussed above. Thus at the rate-limiting step of folding, the NTL9 hydrophobic core contains a few specific interactions involving Ile4 that are stabilized by extensive nonspecific collapse.

In the case of FYN, the original interpretation of the kinetic data suggested that the transition state contains a core folding nucleus comprising positions Ile28, Ala39, and Ile50, surrounded by a loosely packed region encompassing Phe4, Ala6, Leu18, Phe20, and Phe26 (Northey et al. 2002a). The two-parameter analysis is in partial agreement with this model. Ile28 and Ala39 have the largest χ_N values of all positions considered here, reflecting their participation in native-like contacts within the FYN transition state. However, in contrast with the previous model, we have shown that the formation of native-like contacts lags behind hydrophobic burial for side chains at these positions. Within the loosely packed region, only positions 4 and 20 have data available for four or more mutations. χ_N is not significant for Phe20, indicating a near absence of specific contacts, while χ_H is significantly larger than zero. Together, these parameters are consistent with purely nonspecific hydrophobic collapse. Data for Phe4 produce a value of $\chi_N = 0.12$ that is much lower than those of Ile28 and Ala39 yet significantly larger than zero, implying that the side chain at this position is involved in weak, native-like interactions in the transition state. Ile50 was originally included in the core folding nucleus due to the large changes in folding rates caused by mutations at this position. However G_T values show a much stronger dependence on G_H ($r_{TH} = 0.90$) than on G_N ($r_{TN} = 0.39$), and χ_H is significantly larger than zero while χ_N is not. This suggests that Ile50

participates in nonspecific hydrophobic collapse but does not form native-like interactions in the transition state. Estimates of side chain burial based on ($\chi_N + \chi_H$) values show that although the interactions of Ile50 are nonspecific, it is more deeply buried at the rate-limiting step of folding than is Phe4, whose weak interactions have more native character. Thus, hydrophobic core interactions in the FYN transition state involve asynchronous formation of native contacts and side chain burial. The adjacent positions, Ala39 and Ile28 shown in Figure 2B, are both deeply buried and involved in specific interactions. Ile50 is significantly buried but does not participate in specific interactions that impose strong steric constraints, whereas Phe4 is more solvent exposed but forms weak, native-like contacts.

These results have implications for the design and analysis of protein-folding experiments and highlight the detailed information that can be obtained by performing multiple substitutions at individual positions (Mok et al. 2001; Northey et al. 2002a,b; Anil et al. 2005; de los Rios et al. 2005). Disruptive mutations are typically avoided in Φ value analyses because of concerns that these changes may alter the folding pathway (Fersht 1999). We have found that just two parameters can account for the effects of multiple mutations on protein folding rates. Many of the mutations considered in the analysis are disruptive, since they increase side chain volume or introduce polar groups at buried positions. The agreement of these folding data with the two-parameter fits suggests that transition-state structures are quite robust with respect to amino acid sequence changes; therefore a wide range of substitutions can be used in folding studies. In fact, disruptive mutations of the hydrophobicity- and volume-increasing type can yield useful information on the interactions of core side chains in the transition state, since they will tend to decrease folding rates at positions that are specifically packed and increase folding rates at positions that are nonspecifically buried. In fact, if only a single mutation is to be made at a hydrophobic core position, then one that increases side chain volume may be preferable, due to the potential for discriminating specific from nonspecific interactions. Care must be taken when interpreting Φ values obtained from such disruptive mutations, since they are more qualitative indicators of the ability of the transition state as a whole to accommodate a different side chain than they are quantitative measures of the number of side chain contacts. Nevertheless, they can provide useful information on whether a side chain is specifically packed in the protein folding transition state.

In conclusion, we have shown that folding kinetic data for multiple mutations at individual hydrophobic core positions can be well accounted for by two parameters describing the dependence of transition-state stability on the destabilization of the native state and the difference

Figure 2. Structures of (A) the N-terminal domain of ribosomal protein L9 (NTL9: PDB accession code 1DIV) (Hoffman et al. 1994) and (B) the SH3 domain from the Fyn tyrosine kinase (FYN: PDB accession code 1SHF) (Noble et al. 1993) generated using the MOLMOL software package (Koradi et al. 1996). Side chains participating in specific native-like interactions in the transition state ($\chi_N > 0$) are colored orange. Side chains participating purely in nonspecific hydrophobic collapse ($\chi_H > 0$, $\chi_N \approx 0$) are colored blue.

in water/octanol transfer free energies between the wild-type and mutant side chains. These parameters can be interpreted in terms of the formation of native-like structure and additional hydrophobic collapse on a per-site basis. The analysis shows that the process of non-specific burial is generally further progressed than the formation of specific interactions in the transition state, although some native-like contacts are present at the rate-limiting step of folding in all three proteins considered here. We find that the folding transition states examined in this study are fairly tolerant of disruptive mutations, and argue that such substitutions can provide important information regarding side chain interactions in the transition state. The two-parameter fits reported here represent a general approach for interpreting folding data for multiple mutations at hydrophobic core positions and provide more detailed information on the structure of the transition state than do existing methods.

Materials and Methods

We analyzed protein folding data for three proteins in which four or more mutations were made at individual hydrophobic core positions: chymotrypsin inhibitor 2 (CI2) [I30V,A,G,T], the N-terminal domain of ribosomal protein L9 from *Bacillus stearothermophilus* (NTL9) [V31L,Tle,Abu,A; I4Nva,V,Abu,A; L6I,Nva,V,Abu,A; and V21I,L,Abu,A], and the SH3 domain from the Fyn tyrosine kinase (FYN) [F4S,A,V,L; F20S,A,V,L; I28S,A,V,L,F; A39S,V,L,F,G; I50A,V,L,F; V55F,L,W,M,A,H,T,G,S]. In the case of NTL9, both naturally and non-naturally occurring amino acids were employed. The abbreviations Tle, Abu, and Nva refer to tert-leucine, amino-butyric acid, and norvaline, respectively. For CI2, G_T and G_N were taken from the tabulated folding data of Itzhaki et al. (1995). NTL9 data were taken from Anil et al. (2005). FYN Val55 data were taken from de los Rios et al. (2005), while data for all other FYN positions were taken from Northey et al. (2002a). Only positions with greater than 75% burial, as calculated using MOLMOL (Koradi et al. 1996), were considered. For FYN and CI2, G_H values were calculated using water/octanol transfer free energy scale of Fauchere and Pliska (1983). The non-naturally occurring amino acids employed in some NTL9 variants are not included in this scale; therefore, all G_H values for this protein were obtained from calculations based on N-acetyl amino acid amides using software from ChemAxon.

The change in transition-state free energy for a substitution to amino acid i was equated to the change in native-state stability and change in hydrophobicity produced by the mutation according to

$$(G_T^{calc})_i = \chi_N(G_N)_i + \chi_H(G_H)_i, \quad (11)$$

where

$$(G_N)_i = (\Delta G_{U \rightarrow N}^{mut})_i - \Delta G_{U \rightarrow N}^{wt} \\ = -RT((\ln(k_F^{mut})_i - \ln(k_U^{mut})_i) - (\ln(k_F^{wt}) - \ln(k_U^{wt}))), \quad (12)$$

$(k_F^{mut})_i$ and $(k_U^{mut})_i$ are the folding and unfolding rates of the mutant protein, k_F^{wt} and k_U^{wt} are the folding and unfolding rates of the wild-type protein, and $(G_H)_i$ is the octanol to water

transfer free energy of the wild-type side chain subtracted from the octanol to water transfer free energy of the substituted side chain i . The parameters χ_N and χ_H were obtained for each hydrophobic core position by linear regression through the origin (Walpole et al. 1998),

$$\begin{bmatrix} \chi_N \\ \chi_H \end{bmatrix} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{B}, \quad (13)$$

$$\mathbf{A} = \begin{bmatrix} (G_N)_{i=1} & (G_H)_{i=1} \\ \vdots & \vdots \\ (G_N)_{i=n} & (G_H)_{i=n} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} (G_T^{\text{exp}})_{i=1} \\ \vdots \\ (G_T^{\text{exp}})_{i=n} \end{bmatrix}, \quad (14)$$

where the subscript i runs from 1 to n , the number of mutations made at a given position, and $(G_T^{\text{exp}})_i$ is the experimentally observed change in transition-state free energy produced by mutation i , calculated according to Equation 3. Confidence intervals for χ_N and χ_H were estimated from Monte Carlo simulations in which linear regression through the origin was repeated on samples randomized according to the uncertainty in G_T^{exp} values (Efron and Tibshirani 1986). One hundred thousand repetitions were performed for each position, where for iteration j ,

$$\begin{bmatrix} \chi_N \\ \chi_H \end{bmatrix}_j = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \begin{bmatrix} (G_T^{calc})_{i=1} + \varepsilon_{j,i=1} \\ \vdots \\ (G_T^{calc})_{i=n} + \varepsilon_{j,i=n} \end{bmatrix}, \quad (15)$$

and $\varepsilon_{j,i}$ is drawn randomly from a normal distribution with a mean of zero and a standard deviation equal to the uncertainty in $(G_T^{\text{exp}})_i$. The 90% confidence region was taken to be the interval with 5% of iterations lying above the upper bound and 5% of iterations lying below the lower bound. Uncertainties in $(G_T^{\text{exp}})_i$ values, σ_i , were set equal to the reported experimental error or calculated from residuals to the best fit regression line according to

$$\sigma_i = \sqrt{\frac{\sum_{i=1}^n ((G_T^{\text{exp}})_i - (G_T^{calc})_i)^2}{n-2}}, \quad (16)$$

selecting the larger of the two estimates for each mutation. All calculations were performed using the MATLAB software package.

Electronic supplemental material

Separate plots of G_T^{exp} as functions of G_N , G_H , and G_T^{calc} for all 11 hydrophobic core positions included in the analysis are included in the supplemental material.

Acknowledgments

We thank L.E. Kay, J.D. Forman-Kay, and A.R. Davidson for critical reading of this manuscript. This work was supported by grants from the Canadian Foundation for Innovation, the Canadian National Science and Engineering Research Council,

and Le Fonds Québécois de la Recherche sur la Nature et les Technologies.

References

- Anil, B., Sato, S., Cho, J.-H., and Raleigh, D.P. 2005. Fine structure analysis of a protein folding transition state; distinguishing between hydrophobic stabilization and specific packing. *J. Mol. Biol.* **354**: 693–705.
- de los Rios, M.A., Daneshi, M., and Plaxco, K.W. 2005. Experimental investigation of the frequency and substitution dependence of negative ϕ -values in two-state proteins. *Biochemistry* **44**: 12160–12167.
- Efron, B. and Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Stat. Sci.* **1**: 54–77.
- Eriksson, A.E., Baase, W.A., Zhang, X.-J., Heinz, D.W., Blaber, M., Baldwin, E.P., and Matthews, B.W. 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* **255**: 178–183.
- Fauchere, J.-L. and Pliska, V. 1983. Hydrophobic parameters π of amino acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**: 369–375.
- Fersht, A. 1999. *Structure and mechanism in protein science*. W.H. Freeman and Company, New York.
- Fersht, A.R., Itzhaki, L.S., Elmasry, N., Matthews, J.M., and Otzen, D.E. 1994. Single versus parallel pathways of protein-folding and fractional formation of structure in the transition-state. *Proc. Natl. Acad. Sci.* **91**: 10426–10429.
- Hansch, H. and Leo, A. 1979. *Substitution constants for correlation analysis in chemistry and biology*. Wiley, New York.
- Hoffman, D.W., Davies, C., Gerchman, S.E., Hycia, J.H., Porter, S.J., White, S.W., and Ramakrishnan, V. 1994. Crystal structure of prokaryotic ribosomal protein L9: A bi-lobed RNA-binding protein. *EMBO J.* **13**: 205–212.
- Itzhaki, L.S., Otzen, D.E., and Fersht, A.R. 1995. 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.
- Koradi, R., Billeter, M., and Wüthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**: 51–55.
- Ludvigsen, S., Shen, H., Kjaer, M., Madsen, J.C., and Poulsen, F.M. 1991. Refinement of the three-dimensional solution structure of barley serine proteinase inhibitor 2 and comparison with the structures in crystals. *J. Mol. Biol.* **222**: 621–635.
- Matouschek, A., Kellis Jr. J.T., Serrano, L., and Fersht, A.R. 1989. Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340**: 122–126.
- Mok, Y.K., Elisseeva, E.L., Davidson, A.R., and Forman-Kay, J.D. 2001. Dramatic stabilization of an SH3 domain by a single substitution: Roles of the folded and unfolded states. *J. Mol. Biol.* **307**: 913–928.
- Noble, M.E., Mussachio, A., Saraste, M., Courtneidge, S.A., and Wierenga, R.K. 1993. Crystal structure of the SH3 domain in human Fyn; comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. *EMBO J.* **12**: 2617–2624.
- Northey, J.G.B., Di Nardo, A.A., and Davidson, A.R. 2002a. Hydrophobic core packing in the SH3 domain folding transition state. *Nat. Struct. Biol.* **9**: 126–130.
- Northey, J.G.B., Maxwell, K.L., and Davidson, A.R. 2002b. Protein folding kinetics beyond the Φ value: Using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state. *J. Mol. Biol.* **320**: 389–402.
- Richards, F. and Lim, W. 1994. An analysis of packing in the protein folding problem. *Q. Rev. Biophys.* **26**: 423–498.
- Walpole, R.E., Myers, R.H., and Myers, S.L. 1998. *Probability and statistics for engineers and scientists*, 6th ed. Prentice Hall, Upper Saddle River, NJ.
- Weikl, T.R. and Dill, K.A. 2007. Transition-states in protein folding kinetics: The structural interpretation of Φ values. *J. Mol. Biol.* **365**: 1578–1586.