Supporting Information

Naganathan et al. 10.1073/pnas.1000988107

SI Text

Methods. ***Simulation of Brønsted Plots.*** The first scenario assumes that there is a distribution of ϕ-values ranging from 0 to 1 (Fig. 1*A* in the main text). This corresponds to the situation wherein the structural information to be extracted from ϕ-values is maximal. The simulation was performed using the experimental $\Delta\Delta G_{eq}$, which were multiplied by a set of 806 random numbers ranging from 0 to 1 according to a flat distribution to simulate the corresponding $\Delta\Delta G^{\ddagger}_{unfold}$ values (red circles in Fig. 1*A*). Experimental error was simulated adding Gaussian noise of 1 kJ/mol (mean of zero and a standard deviation of one) to both $\Delta\Delta G_{eq}$ and $\Delta\Delta G^{\ddagger}_{unfold}$ (blue circles in Fig. 1*A*). The mean ϕ-value and the correlation coefficients are 0.50 and 0.74, and 0.52 and 0.69, for simulations without and with added noise, respectively. The second scenario is simulated assuming that all mutations produce a single ϕ-value of 0.3 (red line in Fig. 1*B*). $\Delta\Delta G^{\ddagger}_{unfold}$ values were thus obtained by multiplying the experimental $\Delta\Delta G_{eq}$ by 0.3. Adding 1 kJ/mol random noise results in a scatter of points around the average value (blue circles in Fig. 1*B*). The resulting correlation coefficient of this dataset is ~0.94.

Jackknife Analysis. To rule out any inherent bias in the experimental data and to estimate the confidence interval of the mean ϕ-value we performed the nonparametric jackknife analysis. In this method, sample subsets of 50 elements (6.2% of the data) were randomly chosen a million times from the original experimental dataset of 806 mutations. The mean ϕ-value ($\langle\phi\rangle$) and the correlation coefficient were calculated for each subsample. The resulting distribution of $\langle\phi\rangle$ is Gaussian with a mean and standard deviation of 0.24 and 0.04, respectively (blue bars in Fig. S1*A*). The corresponding correlation coefficients show a skewed distribution towards higher values peaking at ~0.91 and with a short tail down to ~0.8 (blue bars in Fig. S1*B*). The $\langle\phi\rangle$ from the correlation analysis of each of the 24 protein datasets is shown as green circles. The only protein that has a $\langle\phi\rangle$ greater than 0.4 is α-spectrin SH3, but it is also the dataset with fewest number of mutations (14) and a narrow experimental range in $\Delta\Delta G_{eq}$.

In parallel, we simulated a million statistical samples of 50 elements picked randomly from the simulation of scenario 1 described in the previous paragraph. The distribution of $\langle\phi\rangle$ (slope of the regression line) and correlation coefficient from the 1 million simulated samples are sharp Gaussian-like and with the following parameters: 0.52 ± 0.02 for $\langle\phi\rangle$ (red bars in Fig. S1*A*) and 0.70 ± 0.02 for the correlation coefficient (red bars in Fig. S1*B*). From this analysis it follows that the $\langle\phi\rangle$ observed experimentally is **seven** standard deviations away from the mean behavior expected for a uniform distribution of ϕ-values, i.e., $(0.52 - 0.24)/0.04$ (highlighted by the dashed cyan lines). If we assume that experimental errors follows a normal distribution this analysis gives a probability $<10^{-10}$ that the experimental results arises from a uniform distribution of ϕ-values between 0 and 1. The same conclusions are extracted from the analysis of the correlation coefficient. Statistical bootstrapping with replacement gave similar results.

Brønsted Analysis at Fixed Mutant Stabilities. In the traditional ϕ-value analysis all mutant effects are referenced to water (i.e., absence of chemical denaturants), and thus the stability of each mutant is different. An alternative is to reference all mutant effects to a fixed common stability. In this case the $\Delta\Delta G_{eq}$ and $\Delta\Delta G^{\ddagger}_{unfold}$ values are obtained in reference to the wild-type at the concentration of chemical denaturant at which the particular mutant attains the target stability. For example, for folding-unfolding isostability conditions ($\Delta G_{eq} = 0$) $\Delta\Delta G_{eq}$ and $\Delta\Delta G^{\ddagger}_{unfold}$ values are obtained at the chemical denaturation midpoints of the mutants. Like in the classical analysis, the global ϕ-value (here $\langle\phi^{iso}\rangle$) is obtained from the slope of the linear regression of the $\Delta\Delta G_{eq}$ versus $\Delta\Delta G^{\ddagger}_{unfold}$ plot.

Structural Analysis. The analysis was performed on 1014 globular proteins that were obtained by filtering the 1520 single-domain proteins between 35 and 150 residues with less than 30% sequence identity available in the PDB according to a radius of gyration criterion ($R_g/\sqrt{N} \leq 1.6$). Residues were considered spatial neighbors if they had at least one atomic contact within 0.68 nm. Local neighbors are those at i, $i+3$ or shorter chain distances. Accessible surface areas were calculated with the STRIDE algorithm (1).

Cluster Analysis. The clustering analysis was performed to group mutations according to the structural-packing properties of the mutated sites. This procedure is critical to uncover any signal from the mutations with apparently higher ϕ-values because these mutations correspond to very small perturbations (~3.1 kJ/mol on average). The uncertainty on the individual ϕ-values for such small perturbations is extremely large, producing 95% confidence ranges that extend beyond the full dynamic range of the ϕ-value (0–1). The large experimental error and additional mutant fluctuations are averaged out within each cluster. Clusters were produced algorithmically minimizing the differences in packing environments between possible elements of each cluster using 10000 rounds of the K-means algorithm from Matlab. Residue packing environments were defined according to three Z-scored properties: number of residues with atoms within 0.68 nm, local/total neighbor ratio, and relative accessible surface area. The number of clusters used in the analysis was selected to maximize the number of clusters while ensuring a uniform distribution of elements between clusters. The optimal for the analysis was found to be 10 clusters with 40, 64, 68, 71, 73, 78, 83, 97, 99 and 106 mutants. However, the results do not depend on the number of clusters as shown in Fig. S2 for 5, 8, 15, and 20 clusters.

1. M. Heinig, D. Frishman (2004) STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* 32:W500–W502.
2. F. Chiti et al. (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat Struct Biol* 6:1005–1009.
3. K.F. Fulton, E.R.G. Main, V. Daggett, S.E. Jackson (1999) Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J Mol Biol* 291:445–461.
4. L. Hedberg, M. Oliveberg (2004) Scattered Hammond plots reveal second level of site-specific information in protein folding: phi′ (beta). *Proc Natl Acad Sci USA* 101:7606–7611.
5. Y. Li, R. Gupta, J.H. Cho, D.P. Raleigh (2007) Mutational analysis of the folding transition state of the C-terminal domain of ribosomal protein L9: A protein with an unusual beta-sheet topology. *Biochemistry* 46:1013–1021.
6. J.C. Martinez, L. Serrano (1999) The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat Struct Biol* 6:1010–1016.
7. L.S. Itzhaki, D.E. Otzen, A.R. Fersht (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.
8. D. S. Riddle et al. (1999) Experiment and theory highlight role of native state topology in SH3 folding. *Nat Struct Biol* 6:1016–1024.

9. D.E. Kim, C. Fisher, D. Baker (2000) A breakdown of symmetry in the folding transition state of protein L. *J Mol Biol* 298:971–984.
10. J.G.B. Northey, A.A. Di Nardo, A.R. Davidson (2002) Hydrophobic core packing in the SH3 domain folding transition state. *Nat Struct Biol* 9:126–130.
11. B.B. Kragelund et al. (1999) The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nat Struct Biol* 6:594–601.
12. H.M. Went, S.E. Jackson (2005) Ubiquitin folds through a highly polarized transition state. *Protein Eng Des Sel* 18:229–237.
13. E.L. McCallister, E. Alm, D. Baker (2000) Critical role of beta-hairpin formation in protein G folding. *Nat Struct Biol* 7:669–673.
14. V. Villegas, J.C. Martinez, F.X. Aviles, L. Serrano (1998) Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J Mol Biol* 283:1027–1036.
15. B. Anil, S. Sato, J.H. Cho, D.P. Raleigh (2005) Fine structure analysis of a protein folding transition state; distinguishing between hydrophobic stabilization and specific packing. *J Mol Biol* 354:693–705.
16. R. Guerois, L. Serrano (2000) The SH3-fold family: Experimental evidence and prediction of variations in the folding pathways. *J Mol Biol* 304:967–982.
17. M.M. Garcia-Mira, D. Boehringer, F.X. Schmid (2004) The folding transition state of the cold shock protein is strongly polarized. *J Mol Biol* 339:555–569.
18. C.T. Friel, A.P. Capaldi, S.E. Radford (2003) Structural analysis of the rate-limiting transition states in the folding of lm7 and lm9: Similarities and differences in the folding of homologous proteins. *J Mol Biol* 326:293–305.
19. F.X. Campbell-Valois, S.W. Michnick (2007) The transition state of the ras binding domain of raf is structurally polarized based on phi-values but is energetically diffuse. *J Mol Biol* 365, 1559–1577 (2007).
20. R.A. Chu, W.H. Pei, J. Takei, Y.W. Bai (2002) Relationship between the native-state hydrogen exchange and folding pathways of a four-helix bundle protein. *Biochemistry* 41:7998–8003.
21. K. Teilum et al. (2005) Different secondary structure elements as scaffolds for protein folding transition states of two homologous four-helix bundles. *Proteins* 59:80–90.
22. M. Petrovich, A.L. Jonsson, N. Ferguson, V. Daggett, A.R. Fersht (2006) phi-Analysis at the experimental limits: Mechanism of beta-hairpin formation. *J Mol Biol* 360:865–881.
23. N. Ferguson, T.D. Sharpe, C.M. Johnson, A.R. Fersht (2006) The transition state for folding of a peripheral subunit-binding domain contains robust and ionic-strength dependent characteristics. *J Mol Biol* 356:1237–1247.
24. S. Sato, A.R. Fersht (2007) Searching for multiple folding pathways of a nearly symmetrical protein: Temperature dependent phi-value analysis of the B domain of protein A. *J Mol Biol* 372:254–267.
25. T.D. Sharpe, N. Ferguson, C.M. Johnson, A.R. Fersht (2008) Conservation of transition state structure in fast folding peripheral subunit-binding domains. *J Mol Biol* 383:224–237.

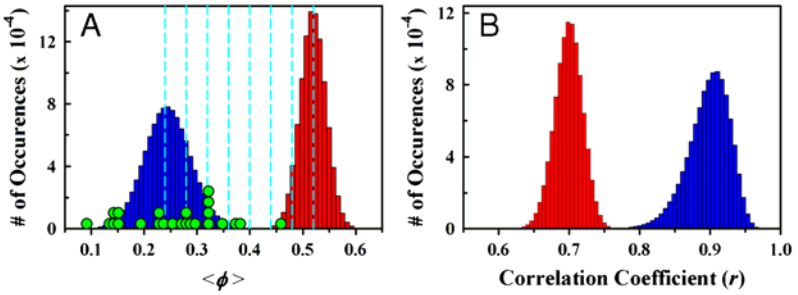

Fig. S1. Results of the jackknife analysis.

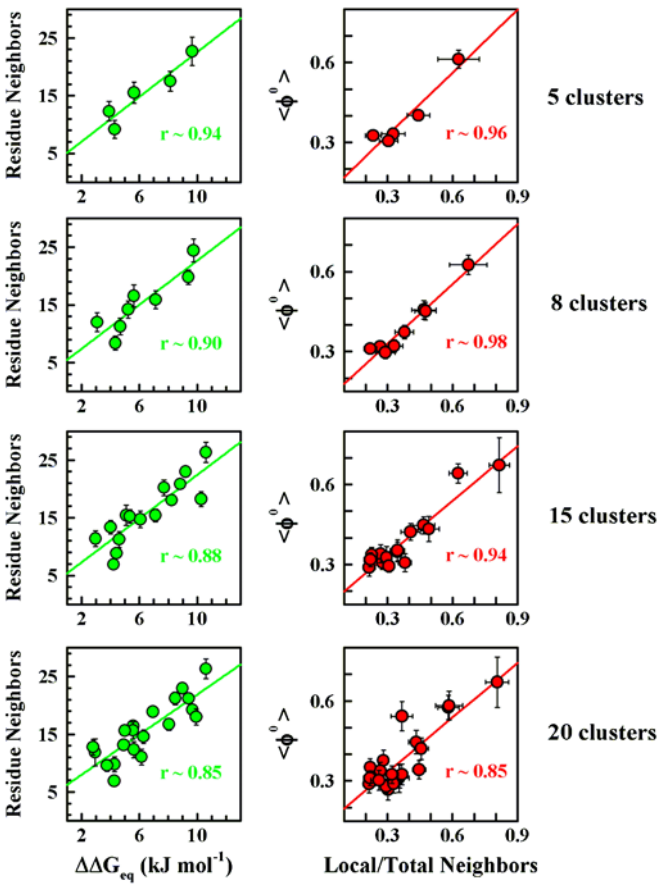

Fig. S2. Results of the cluster analysis for different numbers of clusters. The left column shows the correlation between the number of residue neighbors and the magnitude of the mutant perturbation. The right column shows the correlation between ϕ-value at isostability and the local versus total neighbor ratio.

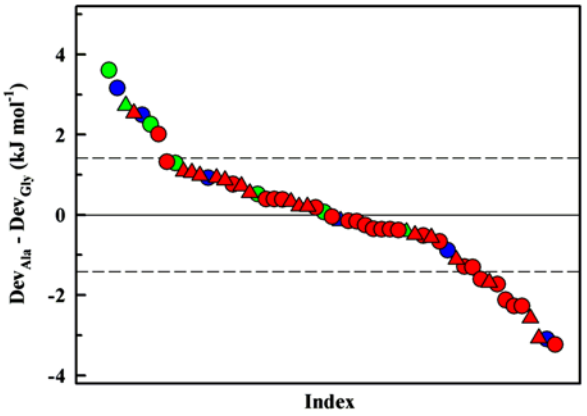

Fig. S3. Comparing the effects of substitutions to Ala or Gly at the same site in conditions of isostability at the chemical denaturation midpoint. The ordinate plots the difference in the deviation from the 0.36 $\Delta\Delta G_{eq}$ line between alanine and glycine. Mutations in α-helical, β-strand and other secondary structure elements are shown in red, blue, and green respectively, with circles and triangles highlighting the buried and exposed positions. To facilitate visual inspection, the mutated sites have been sorted according to the magnitude of the difference in deviation.

Table S1. Proteins studied

Index	Protein	PDB ID	StructureClass	Size	# of Mutations	Reference
1	Muscle AcP	1APS	α/β	98	22	2
2	FKBP12	1FKB	α/β	107	34	3
3	L23	1N88	α/β	96	17	4
4	CTL9	1DIVC	α/β	92	24	5
5	α-spc SH3 D48G	1SHG	β	62	14	6
6	CI2	1COA	α/β	65	65	7
7	src-SH3	1FMK	β	64	54	8
8	Protein L	1HZ6	α/β	62	68	9
9	fyn-SH3	1SHF	β	67	34	10
10	bACBP	2ABD	α	86	30	11
11	Ubq	1UBQ	α/β	76	27	12
12	Protein G	1PGB	α/β	56	31	13
13	ADA2h	1O6X	α/β	81	18	14
14	NTL9	1DIVN	α/β	56	24	15
15	Sso7d	1SSO	β	62	20	16
16	CspB	1CSP	β	67	21	17
17	Im9	1IMQ	α	86	25	18
18	RafRBD	1RFA	α/β	78	47	19
19	Cytb562	1YYJ	α	106	39	20
20	yACBP	1STY	α	86	18	21
21	FBP28 WW	1E0L	β	37	45	22
22	E3BD F166W	1W4E	α	45	22	23
23	BdpA Y15W	1SS1	α	60	45	24
23	BdpA N29HQ33W	1SS1	α	60	20	24
23	BdpA E48W	1SS1	α	60	20	24
24	POB L146AY166W	1W4J	α	51	22	25

Table S2. *m*-value changes

Index	Protein	$\langle \Delta\Delta G_{eq} \rangle$ (kJ/mol)	$\beta_{Tanford}^{WT}$	$\beta_{Tanford}^{Mut}$*	m_{kin}*
1	mAcP	6.36	0.77	1.00	0.99
2	FKBP12	6.39	0.63	1.05	1.13
3	L23	9.91	0.52	1.11	0.93
4	CTL9	10.24	0.69	1.02	1.08
5	α-spc SH3 D48G	6.50	0.71	0.99	1.03
6	CI2	5.34	0.58	1.15	1.17
7	src SH3	3.88	0.65	1.16	1.19
8	protein L	5.61	0.75	1.00	1.15
9	fyn SH3	6.88	0.72	0.96	0.90
10	bACBP	7.12	0.6	1.06	1.03
11	Ubiquitin	11.04	0.66	1.07	1.01
12	protein G	5.21	0.82	0.98	1.15
13	ADA2h	3.85	0.71	1.05	1.05
14	NTL9	5.23	0.69	1.05	1.20
15	sso7d SH3	4.43	0.63	1.05	1.16
16	CspB	4.90	0.88	1.02	1.08
17	Im9	8.38	0.94	0.99	1.03
18	RafRBD	6.09	0.75	1.04	0.99
19	cyt b562	10.32	0.48	1.13	0.91
20	y ACBP	9.97	0.60	1.16	0.96
21	FBP28 WW	2.61	0.71	1.10	1.23
22	E3BD F166W	8.37	0.65	1.12	0.86
23	BdpA Y15W	5.09	0.83	1.04	1.15
24	BdpA N29HQ33W	4.49	0.69	1.06	1.08
25	BdpA E48W	4.87	0.73	1.02	1.02
26	POB L146AY166W	6.74	0.80	1.03	0.94

*The mean value of all the mutations with $|\Delta\Delta G_{eq}| > 5$ kJ/mol referenced to the wild-type value. For FBP28WW domain we set the ΔG_{eq} cutoff to 3 kJ/mol due to the low intrinsic stability of this protein.

Table S3. Mutations that are >2σ away from the mean ϕ^0-value

Index	Protein	Mut.	Deviation (kJ/mol)	$\Delta\Delta G_{eq}$ (kJ/mol)	ϕ^0	Local/Total	m_f*	m_u*	m_{kin}*
1	mAcP	Y11F	−4.67	8.20	0.93	0.15	0.80	0.80	0.80
2	FKBP12	F36A	−6.04	−13.34	−0.09	0.28	1.61	0.62	1.24
3	L23	Y26A	−4.00	−15.69	0.10	0.16	0.87	0.64	0.76
4	CTL9	I93A	−4.38	−14.19	0.05	0.26	0.96	1.15	1.02
5		I98A	−4.71	−20.81	0.13	0.33	1.26	0.94	1.16
6		I115A	−4.51	−15.99	0.08	0.27	1.09	0.88	1.03
7		I121A	−4.02	−13.78	0.07	0.30	1.01	1.00	1.01
8	CI2	A16G	6.58	−8.60	1.12	0.35	1.35	1.00	1.20
9		I30A	4.81	−12.87	0.73	0.44	1.59	0.95	1.32
10		L49A	5.66	−19.34	0.65	0.17	1.54	0.88	1.27
11	src SH3	I34A	6.39	−3.87	2.01	0.33	1.36	0.94	1.22
12		A45G	4.15	−8.38	0.85	0.27	1.67	0.72	1.34
13		S47A	4.08	−9.08	0.81	0.31	1.47	0.81	1.24
14		T50A	4.56	−10.52	0.79	0.46	1.80	0.87	1.48
15		G51A	4.08	−8.55	0.84	0.56	1.48	0.76	1.23
16	protein L	G15A	4.32	−7.64	0.93	0.63	1.20	1.16	1.19
17		A20V	−4.70	7.76	0.96	0.33	0.8	1.08	0.87
18		F62L	−3.96	−12.52	0.04	0.17	1.07	1.22	1.11
19		F62V	−5.48	−14.60	−0.02	0.17	1.00	1.48	1.12
20	fyn SH3	F20S	−4.44	−15.45	0.07	0.23	0.79	0.59	0.73
21		F26I	−4.37	−5.56	−0.43	0.19	0.78	1.11	0.87
22		A39V	−5.00	2.52	2.34	0.25	0.95	1.33	1.06
23		A39F	−5.02	−9.74	−0.16	0.25	0.86	0.77	0.83
24		I50F	−5.32	−14.06	−0.02	0.22	0.84	0.99	0.88
25	bACBP	Q33A	−5.41	−17.10	0.04	0.23	1.55	0.86	1.27
26	Ubiquitin	L43A	−5.20	−20.31	0.10	0.19	1.41	0.50	1.10
27		I61A	−5.28	−13.97	−0.02	0.28	0.83	0.96	0.88
28	protein G	D46A	4.31	−6.90	0.98	0.56	0.97	1.06	0.98
29	NTL9	V3L	−4.53	−10.40	−0.08	0.21	1.27	1.25	1.26
30		V3Tle	−4.73	−9.49	−0.14	0.21	1.31	0.3	1.00
31	CspB	V3T	5.29	−9.07	0.94	0.20	1.40	1.06	1.36
32	Im9	F15A	4.87	−21.14	0.59	0.22	1.03	1.59	1.06
33		F40L	−5.17	−15.53	0.03	0.19	1.02	1.20	1.03
34		L52A	−4.98	−15.86	0.04	0.23	1.03	1.40	1.05
35	RafRBD	V60A	4.16	−9.55	0.79	0.20	1.06	0.90	1.02
36		V70A	4.17	−7.50	0.91	0.28	1.06	0.88	1.01
37	yACBP	L27A	4.43	−9.86	0.81	0.29	1.20	0.91	1.08
38		L80A	5.72	−14.52	0.75	0.29	1.16	1.06	1.12
39	FBP28WW	G16A	5.56	−7.48	1.10	0.83	1.98	1.21	1.76
40	BdpA Y15W	I32G	4.18	−12.97	0.68	0.33	1.37	1.00	1.30

*m_f, m_u and m_{kin} referenced to the wild type. Mutants with a >20% change in either one of these m-values in reference to the wild type are shaded in cyan.