# Characterization of the free-energy landscapes of proteins by NMR-guided metadynamics

Daniele Granata [a], Carlo Camilloni [b], Michele Vendruscolo [b,1], and Alessandro Laio [a,1]

[a] [illegible] (SISSA), [illegible] 34136, [illegible] [b] [illegible] CB2 1EW, [illegible]

[illegible] N. [illegible] 21, 2013 ([illegible] 21, 2012)

The use of free-energy landscapes rationalizes a wide range of aspects of protein behavior by providing a clear illustration of the different states accessible to these molecules, as well as of their populations and pathways of interconversion. The determination of the free-energy landscapes of proteins by computational methods is, however, very challenging as it requires an extensive sampling of their conformational spaces. We describe here a technique to achieve this goal with relatively limited computational resources by incorporating nuclear magnetic resonance (NMR) chemical shifts as collective variables in metadynamics simulations. As in this approach the chemical shifts are not used as structural restraints, the resulting free-energy landscapes correspond to the force fields used in the simulations. We illustrate this approach in the case of the third Ig-binding domain of protein G from streptococcal bacteria (GB3). Our calculations reveal the existence of a folding intermediate of GB3 with nonnative structural elements. Furthermore, the availability of the free-energy landscape enables the folding mechanism of GB3 to be elucidated by analyzing the conformational ensembles corresponding to the native, intermediate, and unfolded states, as well as the transition states between them. Taken together, these results show that, by incorporating experimental data as collective variables in metadynamics simulations, it is possible to enhance the sampling efficiency by two or more orders of magnitude with respect to standard molecular dynamics simulations, and thus to estimate free-energy differences among the different states of a protein with a $k_B T$ accuracy by generating trajectories of just a few microseconds.

NMR [illegible] | [illegible] | [illegible] | [illegible] | [illegible]

In the past two decades, a series of experimental and theoretical advances has made it possible to obtain a detailed understanding of the molecular mechanisms underlying the folding process (1–6). With the increasing power of computers (7), as well as the improvements in force fields (8, 9), atomistic simulations are also becoming increasingly important because they can generate highly detailed descriptions of the motions of proteins (10–12). A supercomputer specifically designed to integrate Newton's equations of motion of proteins (7) recently broke the millisecond time barrier. This achievement has allowed the direct calculation of repeated folding events for several fast-folding proteins (13) and the characterization of molecular mechanisms underlying protein dynamics and function (14). Reliable descriptions of the folding process have also been obtained by exploiting enhanced sampling techniques (15, 16), including replica-exchange molecular dynamics (17), metadynamics (18, 19), and distributed computing (20).

It has also been realized that by bringing together experimental measurements and computational methods, it is possible to expand the range of problems that may be addressed (4, 21–24). For example, by incorporating structural information relative to transition states (TSs; $\phi$ values) as structural restraints in molecular dynamics simulations, it is possible to obtain structural models of these transiently populated states (25, 26), as well as of native (27) and nonnative intermediates (28) explored during the folding process. By applying this strategy to structural parameters measured by NMR spectroscopy, one can determine the atomic-level structures and dynamics of proteins (29–32). In these approaches, the experimental information is exploited to create an additional term in the force field that penalizes the deviations from the measured values, thus restraining the sampling of the conformational space to regions close to those observed experimentally (25).

Here, we propose an alternative strategy to use experimental information to aid molecular dynamics simulations. In this approach, the measured parameters are not used as structural restraints in the simulations but rather to build collective variables (CVs) within metadynamics calculations. In metadynamics (18, 19), the conformational sampling is enhanced by constructing a time-dependent potential that discourages the explorations of regions already visited in terms of specific functions of the atomic coordinates called collective variables. In this work, we show that NMR chemical shifts may be used as collective variables to guide the sampling of conformational space in molecular dynamics simulations.

Because the method that we discuss here enables the conformational sampling to be enhanced without modifying the force field through the introduction of structural restraints, it provides the statistical weights corresponding to the force field used in the molecular dynamics simulations. In the present implementation, we used the bias-exchange metadynamics (BE-META) method (33), an enhanced sampling technique that allows the reconstruction of free energy as a simultaneous function of several variables. By using this approach, we computed the free-energy landscape in explicit solvent of the third Ig-binding domain of streptococcal protein G (GB3). Our calculations predict the native fold as the lowest free-energy minimum, also identifying the presence of an on-pathway compact intermediate with nonnative structural elements. In addition, we provide a detailed atomistic picture of the structure at the folding barrier, which shares with the native state a fraction of the secondary structure elements.

These results have been obtained using relatively limited computational resources. Through the advanced sampling method that we discuss, the total simulation time required to reach convergence in the free energy estimates was 380 ns on seven replicas, which is about three orders of magnitude less than the typical timescale required to fold similar proteins (34). We thus anticipate that the technique introduced here will allow the determination of the free-energy landscapes of a wide range of proteins in cases in which NMR chemical shifts are available.

## Results and Discussion

We performed molecular dynamics simulations of GB3 at 330 K, using the Gromacs 4.5.3 package (35) and the AMBER99SB-ILDN force field (8). To enhance conformational sampling, we used the BE-META scheme (33) with seven replicas. We started the simulations from a structure at 5.7 Å from the reference structure [Protein Data Bank (PDB) ID code 2OED (36)] and ran them for a total of 380 × 7 ns. For each replica, we used

[illegible] D.G., C.C., M.V., [illegible] A.L. [illegible]

[illegible]

[illegible] PNAS [illegible]

[1] [illegible] E-[illegible] 45@[illegible]

[illegible] 10.1073/[illegible] 1218350110/-/DCS[illegible].

a different metadynamics (18) history-dependent potential acting on a different CV (*Methods* and *SI Text*). Three CVs act at the secondary structure level by quantifying, respectively, the fraction of α-helical, antiparallel, and parallel β-sheet content of the protein. Three other CVs act at the tertiary structure level by biasing the number of hydrophobic contacts and the orientation of the side-chain dihedral angles $\chi_1$ and $\chi_2$ for hydrophobic and polar side chains. The seventh CV, called "CamShift," measures the difference between the experimental and calculated chemical shifts, which were obtained using the CamShift method (37) (*Methods* and *SI Text*). Our results indicate that in the approach we present here, this last variable is essential to fold GB3 and reach convergence readily in the free-energy calculations.

**Folding of GB3 Using Chemical Shifts as CVs.** The method that we introduce in this work makes it possible to visit efficiently a wide range of structures, ranging from extended to compact. Representative examples are shown in Fig. 1*A*. Native-like conformations are visited multiple times, reaching a backbone rmsd of 0.5 Å from the reference structure (PDB ID code 2OED). In these native-like structures, the internal packing of hydrophobic side chains is practically identical to that observed in the reference structure (Fig. 1*C*). In the calculations that we performed, this level of accuracy could be reached only by using a bias-exchange scheme in which the CamShift CV is included in the CV set (*Methods* and *SI Text*). To demonstrate this point, we performed another simulation with the same setup, using the six CVs discussed above that describe the secondary and tertiary structures, but not the CamShift CV. The difference between the two simulations is substantial. In the simulation without the CamShift CV, the closest configuration to the reference structure has an rmsd of 2.7 Å (Fig. 2, *Inset B*). After 50 ns, the rmsd starts increasing progressively (red line) and the folded state is not explored at all. By contrast, the simulation with the CamShift CV visits the folded state several times, with several unfolding–refolding events. During the first 50 ns, the latter simulation not only performed better, reaching an rmsd of 2.5 Å, but it also formed the correct secondary and tertiary contacts, particularly the ones involved in forming the first β-hairpin (Fig. 2, *Inset A*), which is critical for the folding of this protein (38, 39). The fraction of native contacts also was systematically higher in the simulation using the CamShift CV (Fig. 2, *Inset C*). These results indicate that the folding events observed later in the simulation are a result of the systematic bias induced by the CamShift CV toward the correct local topology in the native state.

**Thermodynamics of GB3 Folding.** The molecular dynamics simulations that we performed using the approach presented in this work reached convergence after ~240 ns, because at this point the bias potentials acting on all the replicas started to become stationary (40). We then continued the simulations for another 140 ns to reconstruct the free-energy landscape of the protein (*Methods*). In Fig. 3*A*, the free-energy landscape is represented as a function of three CVs: the fraction of antiparallel β-sheet, the fraction of parallel β-sheet, and the coordination number between the hydrophobic side chains (Fig. 3*A*). This representation reveals the organization of the free-energy landscape, with a deep minimum corresponding to native-like structures, separated by a relatively high barrier from other minima. The lowest free-energy minimum (*Methods*) includes configurations very similar to those of the reference structure (on average, at 1.3 Å rmsd).

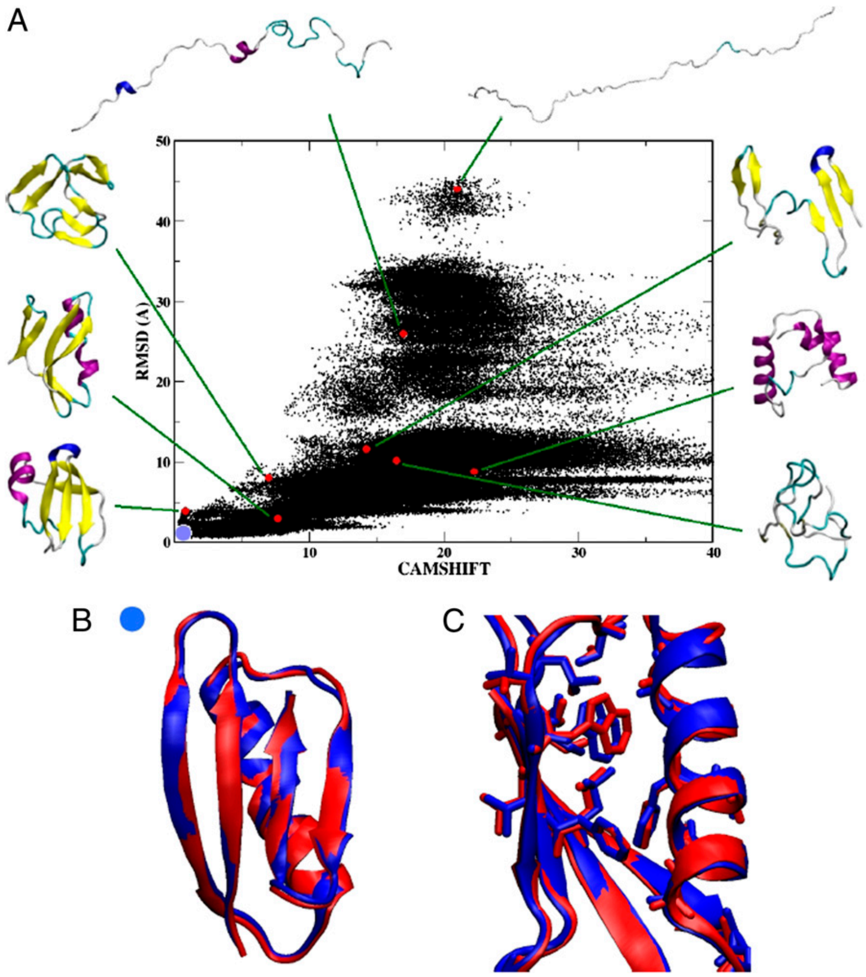

Fig. 1. (A) R[illegible] [illegible] (CV) [illegible] (PDB ID [illegible] 2OED). ( B) S[illegible] (0.5 Å) [illegible] ( C) D[illegible] B.

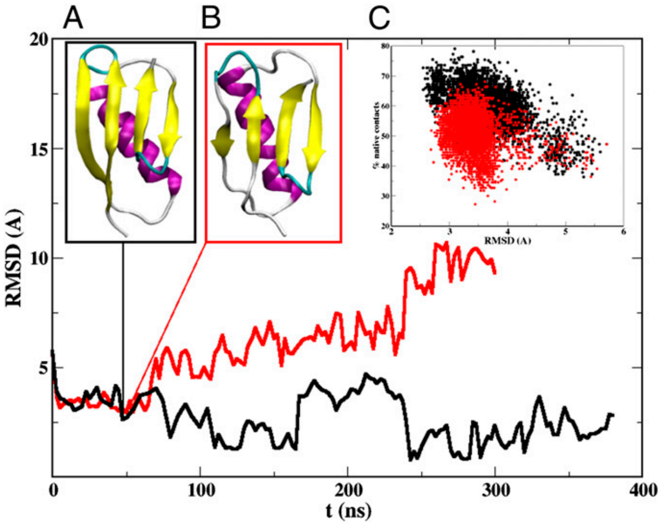

Fig. 2. T[illegible] [illegible](b[illegible]) a[illegible](r[illegible]) [illegible] CS [illegible] CV. (Insets A [illegible] B) L[illegible] ( [illegible] Inset C) P[illegible] [illegible] t50 [illegible] t[illegible]n

This result is confirmed by the analysis of the deviations of the calculated chemical shifts from the corresponding experimental values, both for the reference structure (PDB ID code 2OED) and for the structures belonging to the free-energy minimum (Fig. S1). The agreement is excellent in both cases, thus confirming that by our procedure we could find structures very close to the X-ray structure. These results also provide evidence of the excellent quality of the AMBER99SB-ILDN (8) force field that we used to model GB3.

The shallow minimum immediately after the free-energy barrier separating the folded state from the rest of the conformational space includes compact structures with a high secondary structure content, but with a fold that is rather different from the native, as is discussed below. This second minimum is separated by another free-energy barrier from another minimum, which includes more disordered structures with a much lower secondary structure content. In these conformations, the native C-terminal β-hairpin appears to be present, confirming its high stability, whereas the α-helix and the N-terminal β-hairpin are completely disrupted (41–43). The folded-like and unfolded-like states have a free-energy difference of only 2.3 kJ/mol, which is comparable with the error of our free-energy estimates (40) (*Methods*). The relatively small difference in the free energies of the folded and unfolded states reflects the conformational properties of the protein at the temperature at which the simulation was performed (330 K), which is about 30 K below the experimental melting temperature of GB3 (34).

**An Intermediate State in the Folding of GB3.** The free-energy landscape that we calculated illustrates explicitly the presence of three distinct states of GB3. In addition to the native (N, in dark green in Fig. 3*A*) and unfolded (U, in yellow in Fig. 3*A*) states, we identified the presence of an intermediate state (I, in red in Fig. 3*A*) with a free energy 3.8 kJ/mol higher than that of the N state. From the relative free energies, we calculated the populations of the three states at 330 K, which are 59% for N, 14% for I, and 26% for U. A control unbiased molecular dynamics simulation of 200 ns starting from a structure corresponding to the intermediate free-energy minimum remained extremely stable, with an average rmsd of 2.4 Å from the equilibrated initial structure. These results are consistent with the observation of the presence of an intermediate state of GB1 (44, 45), which shares 88% of the sequence identity of GB3. In particular, that work, which was based on the measurement of the kinetic folding constant as a function of the pH and denaturant concentration, reported a folding behavior consistent with the presence of an on-pathway intermediate and two different TSs (44, 45). However, the structure of the intermediate of GB1 is likely to be more native-like than the one that we find here. The ensemble of conformations making up the intermediate state characterized by our approach contains compact structures, which share specific secondary elements with the native state, including the C-terminal β-hairpin. The N-terminal extension is instead less structured, with only an incipient parallel pairing of the first β-strand (44) and the N-terminal region of the α-helix (residues 22–30). In addition, the C-terminal part of the α-helix exhibits a nonnative configuration by forming an antiparallel β-strand paired with the third β-strand of the protein (residues 41–47).

**Identification and Characterization of the TSs.** To better characterize the folding mechanism of GB3, we simulated by a kinetic Monte Carlo approach (46) the dynamics on the multidimensional free-energy landscape reconstructed by our procedure (*Methods*). All the trajectories connecting the folded and unfolded states go through the intermediate state, confirming that it is an on-pathway intermediate, like the one observed for GB1 (45). The black dashed line in Fig. 3*A* represents the 3D projection of the trajectory of highest probability connecting the folded and unfolded states. Consistent with this topology, the trajectory crosses two TSs: TS1 between the unfolded and intermediate states (in cyan in Fig. 3*A*) and TS2 between the intermediate and native states (in blue in Fig. 3*A*). The rate-limiting step is represented by TS2, with a barrier of 19.5 kJ/mol from the native state, whereas TS1 is at a free energy of 12 kJ/mol.

The hydrophobic solvent-accessible surface area (SASA) reveals how the two TSs are less compact than the N and I states but still quite structured (Fig. 3*B*). A similar conclusion was reached by the experimental Tanford β-values for the two transition states of GB1: $\beta_{TS1} = 0.76 \pm 0.04$ and $\beta_{TS2} = 0.93 \pm 0.04$ (45). These values are consistent with those computed by the ratio of the total SASA between N and the corresponding TS obtained in the present study for GB3, $\beta_{TS1} = 0.82 \pm 0.03$, and $\beta_{TS2} = 0.91 \pm 0.03$.

We found that TS2 of GB3 is more compact than TS1 (Fig. 3*A*), at least in part because of the presence of a native salt bridge between Lys-10 and Glu-56 that is missing in TS1. This aspect also was suggested in the case of GB1 (45) to explain the differences in the pH dependence for the unfolding rate constant of the two TSs. Indeed, an inspection of the TS1, I, and TS2 structures reveals how this salt bridge may trigger the correct arrangement between the C terminus and the first β-strand (residues 1–10). The formation of the salt bridge, which is absent in TS1, acts in I as an anchor that may allow the parallel pairing of the first β-strand, increasing the fraction of native contacts from 29% in I to 37% in TS2. On this view, the second β-hairpin represents the initial native element in the folding process, followed by the formation of the N terminus of the native helix and the parallel pairing of the first β-strand to the C terminus β-hairpin, to then stabilize the formation of the first β-hairpin.

These findings are consistent with the ϕ-values measured for GB1 (38). A comparison between the experimental ϕ-values of GB1 and those calculated for GB3 for TS1 and TS2 is presented in Fig. 4 through the fraction of native contacts of amino acid side chains (25, 26). Despite the differences in sequence between GB1 and GB3, the structure of the TS2 of GB3 exhibits a pattern approximately consistent with experimental ϕ-values of the TS of GB1 (Fig. 4), especially in the two β-hairpin regions. These results, which are consistent with previous conclusions (38), indicate that in the TS the C-terminal hairpin is completely formed as well as the parallel pairing of the first β-strand. Instead, the ϕ-values in the α-helical region show a more complex behavior compatible with a variety of conformations in the transition ensemble.

## Conclusions

We have introduced a method for calculating the free-energy landscapes of proteins based on the incorporation of experimental

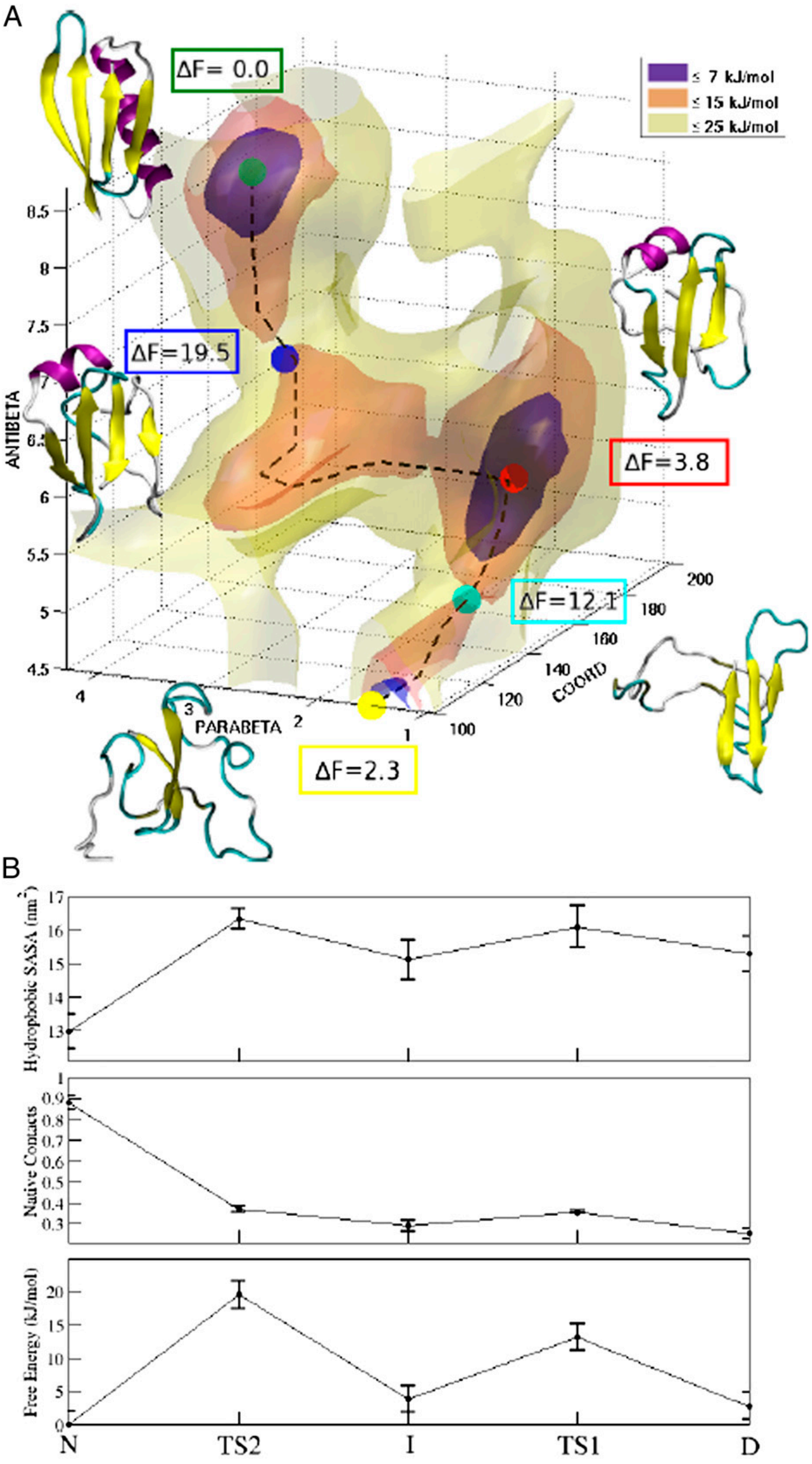

Fig. 3. (A) T[illegible] [illegible] GB3 [illegible] CVs ([illegible]) A[illegible] [illegible] N [illegible] TS2 [illegible] TS1 [illegible] (B) H[illegible] SASA, [illegible]

NMR chemical shifts as collective variables in bias-exchange metadynamics simulations (18, 33). To this end, we have defined a collective variable that measures the difference between experimental and calculated chemical shifts, and helps the simulations find a route to the folded state by exploiting the capability of the chemical shifts to characterize in detail the local configuration of a protein molecule. We have found that this procedure facilitates the formation of the correct native contacts and, consequently, the identification of structures effectively indistinguishable from the native-state conformation determined experimentally.

A distinctive aspect of the approach that we have presented is that it uses the chemical shifts only to define a reaction co-

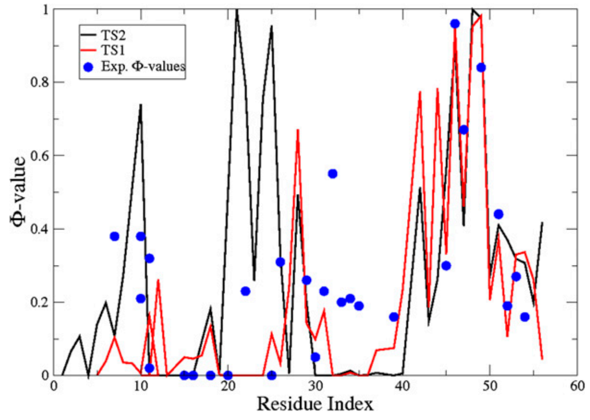

Fig. 4. C[illegible] ϕ-[illegible] GB1 (38) [illegible] ϕ-[illegible] GB3 [illegible] TS2 ([illegible] TS1 ([illegible]

ordinate, without modifying the underlying forcefield used in the molecular dynamics simulations. Hence, the resulting free-energy landscape derives from the Boltzmann distribution of the system for the forcefield used in the simulations.

This procedure allows the free-energy landscapes of proteins to be determined with relatively limited computational resources. In the case of GB3 discussed here, only 380-ns simulations for seven replicas were required. Our calculations have revealed the presence of an on-pathway, partially nonnative intermediate state and have enabled us to estimate accurately the free-energy differences between the different states populated by this protein.

Because the approach that we have described is based on the use of chemical shifts as collective variables in metadynamics simulations, it also may be adopted if only incomplete data are available. It thus will be interesting to explore its applicability to larger proteins and to intrinsically disordered proteins (IDPs) to generate ensembles of structures and free-energy landscapes consistent with chemical shift data. Furthermore, this kind of approach may be generalized by incorporating other experimental data in a metadynamics framework, including NOEs, J-couplings, and residual dipolar couplings, or data from other experimental techniques, such as Small-Angle X-ray Scattering (SAXS) or Fluorescence Resonance Energy Transfer (FRET) methods. We anticipate that these developments will provide molecular dynamics descriptions of the behavior of a variety of proteins for which only sparse experimental data are available.

## Methods

**BE-META.** B[illegible] (BE-META) [illegible] (17) [illegible] (18), [illegible] CV [illegible] CV ([SI Text](#)). E[illegible] A [illegible] (33). T[illegible] CVs [illegible] CVs [illegible] H[illegible] (47), [illegible] CV [illegible] NMR [illegible] (30) [illegible]

**CamShift CV.** T[illegible] NMR [illegible] CamShift (37), [illegible] ([SI Text](#)). U[illegible] (48–50), [illegible] CamShift [illegible] CV [illegible] ($^1H_{\alpha}$, $^{13}C_{\alpha}$, $^{13}C_{\beta}$, $^{13}C'$, $^1H_N$, $^{15}N$) ([SI Text](#) [illegible] Fig. S2) (30, 32). B[illegible] CV [illegible] E[illegible] (37). T[illegible] CV [illegible] T[illegible] CamShift CV [illegible] PLUMED (51) [illegible] G[illegible] (35).

**Simulation Details.** We [illegible] BE-META [illegible] GB3 [illegible] 330 K, [illegible] CVs ([SI Text](#)):

- C[illegible] T[illegible] CV [illegible] P[illegible] G[illegible] $\sigma = 1$.
- A[illegible] RMSD, P[illegible] RMSD, [illegible] A[illegible] RMSD: T[illegible] CVs [illegible] α-[illegible] β-[illegible] (47). P[illegible] A[illegible] RMSD, $m = 4$, $n = 2$, $R_0 = 0.08$, [illegible] $\sigma = 0.2$; [illegible] P[illegible] RMSD, $m = 12$, $n = 8$, $R_0 = 0.08$, [illegible] $\sigma = 0.1$; [illegible] A[illegible] RMSD, $m = 12$, $n = 8$, $R_0 = 0.08$, [illegible] $\sigma = 0.2$.
- C[illegible] N[illegible] T[illegible] CV [illegible] P[illegible] $m = 8$, $n = 4$, $R_0 = 0.4$, [illegible] $\sigma = 10$.
- T[illegible] A[illegible] S[illegible] T[illegible] CVs [illegible] $\chi_1$ [illegible] $\chi_2$, [illegible] P[illegible] $\sigma = 0.5$ [illegible]

T[illegible] CVs [illegible] 51 ([SI Text](#)). S[illegible] 5.7 Å [illegible] PDB ID 2OED (36)] [illegible] 380 [illegible] G[illegible] 4.5.3 [illegible] (35) [illegible] AMBER99SB-ILDN [illegible] (8) [illegible] TIP3P [illegible] (52). T[illegible] 6,524 [illegible] 212-m[illegible] $^3$ [illegible] T[illegible] (53) [illegible] 1 m A [illegible] L[illegible] –J[illegible] 1.2 m A [illegible] LINCS (LIN[illegible] C[illegible] S[illegible] (54). T[illegible] 2.0 f[illegible] N[illegible] –H[illegible] (55, 56) [illegible] 1 [illegible] T[illegible] 1 p[illegible] C[illegible] 1D G[illegible] $w = 0.30$ kJ/[illegible] 4 p[illegible] 20 p[illegible]

A[illegible] 120 [illegible] CVs [illegible] (57). A[illegible] 0.5 [illegible] G[illegible] CamShift CV [illegible] $\sigma$ [illegible] A[illegible] RMSD CV ([SI Text](#) [illegible] Table S1). T[illegible] BE-META, [illegible] PLUMED [illegible] (51) [illegible] G[illegible] A[illegible] BE-META [illegible] 300 [illegible] CamShift CV, [illegible] CV [illegible] F[illegible] 200 [illegible]

**Free-Energy Reconstruction in the CV Space.** BE-META [illegible] (33) ([SI Text](#)). T[illegible] $t_{eq} = 240$ [illegible] A[illegible] CVs [illegible] CV [illegible] CV [illegible] ([SI Text](#)). T[illegible] T[illegible] $t_{eq}$. In [illegible] CamShift C[illegible] N[illegible] A[illegible] P[illegible] B[illegible] RMSD CVs, [illegible] Fig. S3. T[illegible]

Fig 3 [illegible] 3 k/bn Al [illegible] (40) ( SI Text ), [illegible] METAGUI (58), a V[illegible] M[illegible] D[illegible] (VMD) (59) [illegible]

ACKNOWLEDGMENTS. W[illegible] (HPC) G[illegible] 2011 [illegible] CASPUR S[illegible] [illegible] R[illegible] C[illegible] 5 [illegible] M[illegible] G[illegible] Rf 12214. C.C. [illegible] M[illegible] C[illegible] E[illegible]

1. F[illegible] H, S[illegible] G, W[illegible] PG (1991) T[illegible] Science 254(5038):1598–1603.
2. F[illegible] A (1998) Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding (F[illegible])
3. M[illegible] V, E[illegible] WA (1999) A [illegible] Proc Natl Acad Sci USA 96(20): 11311–11316.
4. D[illegible] C, S[illegible] A, K[illegible] M (1998) P[illegible] A [illegible] Angew Chem Int Ed 37(7):868–893.
5. W[illegible] PG, O[illegible] N, T[illegible] D (1995) N[illegible] Science 267(5204):1619–1620.
6. D[illegible] KA, C[illegible] HS (1997) F[illegible] Nat Struct Biol 4(1): 10–19.
7. S[illegible] DE, [illegible] (2008) A[illegible] Commun ACM 51(7):91–97.
8. P[illegible] S, L[illegible] K, S[illegible] DE (2011) H[illegible] Biophys J 100(9):L47–L49.
9. M[illegible] AD, Jr, B[illegible] N, F[illegible] N (2000-2001) D[illegible] CHARMM [illegible] Biopolymers 56(4):257–265.
10. K[illegible] M, M[illegible] A (2002) M[illegible] Nat Struct Biol 9(9):646–652.
11. P[illegible] VS, [illegible] (2003) A[illegible] Biopolymers 68(1):91–109.
12. B[illegible] RB (2012) A[illegible] Curr Opin Struct Biol 22(1):52–61.
13. L[illegible] K, P[illegible] S, D[illegible] RO, S[illegible] DE (2011) H[illegible] Science 334(6055):517–520.
14. S[illegible] DE, [illegible] (2010) A[illegible] Science 330(6002):341–346.
15. C[illegible] M, [illegible] WF (2008) O[illegible] A [illegible] J Comput Chem 29(2):157–166.
16. B[illegible] PG, C[illegible] D, D[illegible] C, G[illegible] PL (2002) T[illegible] Annu Rev Phys Chem 53: 291–318.
17. S[illegible] Y O[illegible] (1999) R[illegible] Chem Phys Lett 314(1-2):141–151.
18. L[illegible] A, P[illegible] M (2002) E[illegible] Proc Natl Acad Sci USA 99(20):12562–12566.
19. L[illegible] A, G[illegible] F (2008) M[illegible] A [illegible] Rep Prog Phys 71(12):126601.
20. S[illegible] M, P[illegible] VS (2000) S[illegible] Science 290(5498): 1903–1904.
21. F[illegible] AR, D[illegible] V (2002) P[illegible] Cell 108(4):573–582.
22. S[illegible] M, M[illegible] F, W[illegible] M, H[illegible] T, O[illegible] N (2012) G[illegible] Proc Natl Acad Sci USA 109(26):10340–10345.
23. R[illegible] AH, [illegible] (2010) H[illegible] RNA [illegible] Nature 468(7324):713–716.
24. V[illegible] M[illegible] U, [illegible] (2012) A[illegible] Proc Natl Acad Sci USA 109(46):18821–18826.
25. V[illegible] M, P[illegible] E, D[illegible] CM, K[illegible] M (2001) T[illegible] Nature 409(6820):641–645.
26. P[illegible] E, V[illegible] M, D[illegible] CM, K[illegible] M (2002) D[illegible] J Mol Biol 324(1):151–163.
27. G[illegible], C[illegible] A (2002) M[illegible] Proc Natl Acad Sci USA 99(10):6719–6724.
28. De S[illegible] A, M[illegible] RW, V[illegible] M (2011) D[illegible] J Chem Theory Comput 7(12):4189–4195.
29. L[illegible] K, B[illegible] RB, D[illegible] MA, D[illegible] CM, V[illegible] M (2005) S[illegible] Nature 433(7022):128–132.
30. R[illegible] P, K[illegible] K, C[illegible] A, V[illegible] M (2010) U[illegible] NMR [illegible] Structure 18(8): 923–933.
31. N[illegible] P, [illegible] (2012) S[illegible] Science 336(6079):362–366.
32. C[illegible] C, R[illegible] P, De S[illegible] A, C[illegible] A, V[illegible] M (2012) C[illegible] RN[illegible] A [illegible] NMR [illegible] J Am Chem Soc 134(9):3968–3971.
33. P[illegible] S, L[illegible] A (2007) A [illegible] J Phys Chem B 111(17):4553–4559.
34. A[illegible] P, O[illegible] , B[illegible] P (1992) K[illegible] 56 [illegible] G-[illegible] G. Biochemistry 31(32): 7243–7248.
35. H[illegible] B, K[illegible] C, V[illegible] D S[illegible] D, L[illegible] E (2008) G[illegible] A: A[illegible] J Chem Theory Comput 4(3):435–447.
36. U[illegible] TS, R[illegible] BE, D[illegible] F, B[illegible] A (2003) E[illegible] NMR [illegible] J Am Chem Soc 125(30):9179–9191.
37. K[illegible] KJ, R[illegible] P, C[illegible] A, S[illegible] X V[illegible] M (2009) F[illegible] NMR [illegible] J Am Chem Soc 131(39):13894–13895.
38. M[illegible] EL, A[illegible] , B[illegible] D (2000) C[illegible] G [illegible] Nat Struct Biol 7(8):669–673.
39. C[illegible] C, B[illegible] RA, T[illegible] G (2011) H[illegible] G, CI2, [illegible] ACBP [illegible] J Chem Phys 134(4): 045105.
40. M[illegible] F, P[illegible] F, L[illegible] A, P[illegible] S (2009) A [illegible] PLOS Comput Biol 5(8):e1000452.
41. B[illegible] FJ, S[illegible] L (1995) F[illegible] G B1 [illegible] Eur J Biochem 230(2):634–649.
42. B[illegible] G, L[illegible] A, P[illegible] M (2006) E[illegible] Phys Rev Lett 96(9):090601.
43. C[illegible] C, P[illegible] D, T[illegible] G, B[illegible] RA (2008) E[illegible] G [illegible] Proteins 71(4):1647–1654.
44. P[illegible] SH, O’N[illegible] KT, R[illegible] H (1997) A[illegible] B1 [illegible] G [illegible] Biochemistry 36(47): 14277–14283.
45. M[illegible] A, [illegible] (2011) GB1 [illegible] Biophys J 101(8):2053–2060.
46. B[illegible] A, K[illegible] M, L[illegible] (1975) A [illegible] J Comput Phys 17(1):10–18.
47. P[illegible] F, L[illegible] A (2009) A [illegible] A[illegible] B [illegible] J Chem Theory Comput 5(9): 2197–2201.
48. H[illegible] B, L[illegible] Y G[illegible] SW, W[illegible] DS (2011) SHIFTX2: S[illegible] J Biomol NMR 50(1):43–57.
49. S[illegible] Y D[illegible] F, C[illegible] G, B[illegible] A (2009) TALOS+: A [illegible] NMR [illegible] J Biomol NMR 44(4):213–223.
50. S[illegible] Y B[illegible] A (2010) SPARTA+: A [illegible] NMR [illegible] J Biomol NMR 48(1):13–22.
51. B[illegible] M, [illegible] (2009) P[illegible] A [illegible] Comput Phys Commun 180(10):1961–1972.
52. J[illegible] W, C[illegible] J, M[illegible] J, I[illegible] R, K[illegible] M (1983) C[illegible] J Chem Phys 79:926–935.
53. E[illegible] U, [illegible] (1995) A [illegible] J Chem Phys 103: 8577–8593.
54. H[illegible] B, [illegible] (1997) L[illegible] A [illegible] J Comput Chem 18(12):1463–1472.
55. N[illegible] S (1984) A [illegible] Mol Phys 52(2):255–268.
56. H[illegible] WG (1985) C[illegible] E[illegible] Phys Rev A 31(3):1695–1697.
57. B[illegible] F, C[illegible] P, P[illegible] F, L[illegible] A (2012) P[illegible] Curr Phys Chem 2:79–91.
58. B[illegible] X, P[illegible] F, M[illegible] F, L[illegible] A (2012) M[illegible] A [illegible] Comput Phys Commun 183(1): 203–211.
59. H[illegible] W, D[illegible] A, S[illegible] K (1996) VMD: V[illegible] J Mol Graph 14(1):33–38, 27–28.

# Supporting Information

**Granata et al. 10.1073/pnas.1218350110**

## SI Text

### 1. Metadynamics

Metadynamics is a computational technique aimed at enhancing the sampling of the conformational space of complex molecular systems (1). The enhancement is obtained through a bias that acts on a small number of parameters, referred to as collective variables (CVs), $s(x)$, which provide a coarse-grained description of the system and are explicit functions of the Cartesian coordinates $x$. The bias takes the form of a history-dependent potential constructed as a sum of Gaussian distributions centered along the trajectory of the CVs (2):

$$V_G(s(x),t) = w \sum_{t'=\tau_G, 2\tau_G, \ldots} exp\left(-\frac{\left(s(x) - s\left(x(t')\right)\right)^2}{2\sigma_s^2}\right), \qquad \textbf{[S1]}$$

where the sum is taken for $t' < t$. Three parameters enter into the definition of $V_G$: (*i*) the height $w$ of the Gaussian distributions, (*ii*) the width $\sigma_s$ of the Gaussian distributions, and (*iii*) the frequency $\tau_G^{-1}$ at which the Gaussian distributions are deposited.

These three parameters influence the accuracy and efficiency of the free-energy reconstruction. If the Gaussian distributions are large, the free-energy surface will be explored at a fast pace, but the reconstructed profile will be affected by large errors. If instead the Gaussian distributions are small or are deposited infrequently, the reconstruction will be accurate, but it will take longer. Typically, the width $\sigma_s$ is chosen to be of the order of the standard deviation of the CV in a preliminary unbiased simulation in which the system explores a local minimum on the free-energy surface (2). In time, the bias potential fills the minima on the free-energy surface, allowing the system to efficiently explore the space defined by the CVs. It is possible to show that in the limit of long times, $V_G(s, t)$  $-F(s)$ (3). Qualitatively, as long as the CVs are uncorrelated, the time required to reconstruct a free-energy surface for a given accuracy scales exponentially with the number of CVs. Therefore, the performance of the algorithm rapidly deteriorates as the dimensionality of the CV space increases. This aspect makes it impractical to obtain an accurate calculation of the free energy when the dimensionality of the space is large. Unfortunately, this often is the case for complex reactions such as protein folding, in which it is very difficult to select a priori a limited number of variables that describe the process, at least unless the structure of the native state is not taken into account explicitly in the CVs.

### 2. Bias-Exchange Metadynamics

The bias-exchange metadynamics (BE-META) method was proposed to overcome the difficulties discussed above (4). The BE-META method involves a combination of replica exchange (5) and metadynamics, in which multiple metadynamics simulations of the system at a given temperature are performed. Each replica is biased with a time-dependent potential acting on a different CV. Exchanges between the bias potentials in the different variables are allowed periodically according to a replica-exchange scheme. If the exchange move is accepted, the trajectory that previously was biased in the direction of the first variable continues its evolution biased by the second, and vice versa. In this manner, a relatively large number of different variables can be biased, and a high-dimensional space may be explored after a sufficient number of exchanges. The result of the simulation, however, is not a free-energy hypersurface in several dimensions, but several (less informative) low-dimensional projections of the free-energy surface along each of the CVs. The high-dimensional hypersurface still can be reconstructed (6) using the method summarized in in *SI Text*, section 5.

### 3. Choice and Definition of CVs in the BE-META Method

Similar to other methods that reconstruct the free energy as a function of a set of generalized coordinates, in the BE-META method the choice of CVs plays an essential role in determining the convergence and efficiency of the free-energy calculation. If the chosen set of CVs does not distinguish different metastable states of the system, the simulation will be affected by hysteresis because not all the important regions of the conformational space will be explored. To choose an appropriate set, one needs to exploit some basic knowledge on the topological, chemical, and physical properties of the system. Although there is no a priori recipe for finding the correct set of CVs, in the BE-META method the number of variables may be relatively large, making the selection less critical.

To study the free-energy landscape of GB3, we used the following CVs:

- AlphaRMSD, ParaBetaRMSD, and AntiBetaRMSD: These CVs count how many fragments of six residues (six in a row for α-helices and three plus three for β-sheets) belong to an α-helix and β-sheet, by computing their rmsd with respect to an ideal α-helix and β-sheet conformation (7):

  $$S = \sum_{\alpha} n\left[\mathrm{rmsd}\left(\{\mathbf{R}_i\}_{i\in\Omega_\alpha}, \{\mathbf{R}^0\}\right)\right] \qquad \textbf{[S2]}$$

  $$n(\mathrm{rmsd}) = \frac{1-(\mathrm{rmsd}/0.1)^n}{1-(\mathrm{rmsd}/0.1)^m}, \qquad \textbf{[S3]}$$

  where $n$ is a function switching smoothly between 0 and 1, the rmsd is measured in nanometers, and $\{\mathbf{R}_i\}_{i\in\Omega\alpha}$ are the atomic coordinates of a set $\Omega_\alpha$ of six residues of the protein, whereas $\{\mathbf{R}^0\}$ are the corresponding atomic positions of ideal α-helical and β-sheet conformations; $m$, $n$ are exponents that allow tuning of the smoothness of the function.

- Coordination Number: This CV, which is used to quantify the number of contacts between the side chain heavy atoms of hydrophobic residues, is defined as

  $$C_N = \sum_{i,j} C_{ij}$$

  with

  $$C_{ij} = \frac{1-\left(\dfrac{r_{ij}}{r_0}\right)^n}{1-\left(\dfrac{r_{ij}}{r_0}\right)^m}, \qquad \textbf{[S4]}$$

  where $r_{ij}$ is the distance between atoms or groups $i$ and $j$, $r_0$ is the distance value to consider two atoms in contact, and $m$, $n$ are exponents that allow tuning of the smoothness of the function.

- AlphaBeta Similarity: We considered two CVs of this type, corresponding to the $\chi_1$ and $\chi_2$ side-chain dihedral angles, respectively, for hydrophobic and polar amino acids. These

CVs are designed to enhance the side-chain packing searching, which is crucial for protein folding. The CVs are defined as

$$AB_{Sim} = \sum_i \frac{1}{2}\left[1 + cos\left(\chi_i - \chi_i^{ref}\right)\right], \qquad \textbf{[S5]}$$

where we have chosen $\chi_i^{ref}$ as the mean value of the corresponding dihedral angle from a library of folded proteins extracted from the Protein Data Bank (PDB).

- CamShift: This CV is used to enhance the conformational sampling along a variable that takes into account the difference between experimental and calculated chemical shifts. Its implementation is described in in *SI Text*, section 4.

All these CVs are implemented in PLUMED (8) for Gromacs (9). They are available publicly and, except for CamShift, were used in several previous protein-folding studies by metadynamics (6, 7).

**Choice of Parameters.** As mentioned in *SI Text*, section 1, the choice of parameters $w$ and $\sigma_s$ influences the accuracy and efficiency of the free-energy reconstruction. Artifacts tend to arise when the free-energy landscape is highly inhomogeneous, being characterized by the simultaneous presence of very shallow and very narrow free-energy basins (10). The parameters of the Gaussian distributions should be chosen so that the maximum force introduced by a single Gaussian distribution ($w/\sigma_s$) is smaller than the typical derivative of the free energy (10). To choose these parameters, we follow a previously proposed scheme (2). In particular, the width $\sigma_s$ was chosen to be of the order of the SD of the CV, performing several preliminary unbiased simulations starting from different folded and unfolded configurations, in which the system explored a local minimum in the free-energy surface. Following this procedure, we verified that the choice of parameters was correct (Table S1). Indeed, the force introduced by a single Gaussian distribution is smaller than the typical derivative of the free energy for the different CVs. This result is also confirmed by the shape of the free-energy projections (Fig. S3) along the CVs used in the analysis (in *SI Text*, section 5): the profiles are homogeneous and smooth, with the minima wider than the Gaussian width. All values of the parameters used in this work are reported in *Simulation Details* in the main text.

## 4. Implementation of the CamShift CV

The implementation of CamShift as a CV requires the structure-based calculations of the chemical shifts. In this work, the chemical shift of a given atom is calculated as (11)

$$\begin{aligned}\delta_{calc} = \delta_{coil} + \delta_{dihedrals} + \delta_{rings} + \delta_{backbone} \\ + \delta_{side-chains} + \delta_{through-space},\end{aligned} \qquad \textbf{[S6]}$$

where, $\delta_{coil}$ is a residue-dependent constant and $\delta_{dihedrals}$ is calculated using the ϕ, ψ, and $\chi_1$ dihedral angles as

$$\delta_{dihedrals} = p_1\cos(3(\theta + p_4)) + p_2\cos(\theta + p_5) + p_3, \qquad \textbf{[S7]}$$

where $p_i$ are given coefficients. The $\delta_{rings}$ term, which takes into account the ring current contributions, is defined using the classical point-dipole method (12). The $\delta_{backbone}$, $\delta_{side-chains}$, and $\delta_{through-space}$ terms are defined as

$$\delta_X = \sum_{j,k} \alpha_{jk}\, d_{jk}^{\beta_{jk}}, \qquad \textbf{[S8]}$$

where $j,k$ defines a pair of atoms at distance $d$; α and β are given coefficients. For $\delta_{backbone}$, the atoms are selected from the neighboring residues along the chain; for $\delta_{side-chains}$, the atoms are those of the same residue; whereas, for $\delta_{through-space}$, the atoms are selected among those within a radius of 0.5 nm and do not belong to the current and neighboring residues.

Because all these terms are defined as differentiable functions of the atomic coordinates, it is possible to compute their derivatives and the corresponding forces in molecular dynamics simulations (13, 14). The collective variable then is defined as

$$CamShift(t) = \sum_{i=1}^{N}\sum_j E_{ij}, \qquad \textbf{[S9]}$$

where $i$ runs over the residues of the protein and $j$ runs over the different atom types ($H_\alpha$, $H_N$, N, $C_\alpha$, $C_\beta$, and C′). $E_{ij}$ has the functional form (13, 14):

$$E_{ij} = \begin{cases} 0 & \text{if} \quad \left|\delta_{calc}^{ij} - \delta_{exp}^{ij}\right| \le n\epsilon_j \\ \left(\dfrac{\left|\delta_{calc}^{ij} - \delta_{exp}^{ij}\right| - n\epsilon_j}{\beta_j}\right)^2 & \text{if} \quad n\epsilon_j < \left|\delta_{calc}^{ij} - \delta_{exp}^{ij}\right| \le x_0 \\ \left(\dfrac{x_0 - n\epsilon_j}{\beta_j}\right)^2 + \gamma\tanh\left(\dfrac{2(x_0 - n\epsilon_j)\left(\left|\delta_{calc}^{ij} - \delta_{exp}^{ij}\right| - x_0\right)}{\gamma\beta_j^2}\right) & \text{for} \quad x_0 < \left|\delta_{calc}^{ij} - \delta_{exp}^{ij}\right|, \end{cases} \qquad \textbf{(S10)}$$

where $\delta_{exp}$ and $\delta_{calc}$ are the experimental and calculated chemical shifts, respectively. The function $E_{ij}$ has a flat bottom (Fig. S2) so that the chemical shifts calculated to within a given accuracy of the experimental value do not produce a penalty. The width of the flat region of the potential is determined by the term $n\varepsilon_j$, where $n$ is a tolerance parameter and $\varepsilon_j$ is the accuracy of the CamShift predictions used for the chemical shifts of type $j$ (11). The penalty is harmonic until the deviation reaches a cutoff value $x_0$, at which point the penalty grows according to a hyperbolic tangent function defined to maintain a continuous derivative at the point $x_0$. The magnitude of the penalty is scaled for each chemical shift type $j$ by the variable $\beta_j$, which is a function of the variability of that chemical shift in folded proteins reported in the Biological Magnetic Resonance Bank database (15). The scaling factor $\beta_j$ is used to obtain relative contributions of comparable magnitude of each chemical shift type to the CV value. The parameter γ determines how large the penalty can grow for deviations beyond $x_0$. In this investigation, the simulation was run with $n = 1$ for all chemical shifts. The harmonic truncation point $x_0$ was set to 4.0 ppm for $H_\alpha$ and $H_N$, and 20.0 ppm for N, $C_\alpha$, $C_\beta$, and C′. The penalty truncation factor γ was set to 20 for all chemical shifts. These values of $x_0$ and $\gamma$ result in an essentially harmonic penalty for most chemical shifts, with penalties reaching the hyperbolic tangent region of the penalty function only in the case of very large outliers (13, 14).

The CV and the forces, which were derived analytically, have been implemented explicitly into a modified version of PLUMED (8) that will be made public in a future release.

In Fig. S3, we report the comparison between the structures sampled in the free-energy minimum and the experimental reference (PDB ID code 2OED), showing the difference between the chemical shifts calculated by the CamShift method and the corresponding experimental values (16) for the different atom types.

## 5. Free Energy Reconstruction

The BE-META method allows the free energy of the system to be reconstructed once the bias potentials become stable (4). To estimate the relative probability of the different states, the low-dimensional free-energy surfaces obtained from the BE-META calculations are exploited to estimate, by a weighted-histogram procedure, the free energy of a finite number of structures representative of all the configurations explored by the system. The CV space is subdivided so that all the frames of the BE-META trajectories are grouped in sets (microstates) whose members are close to each other in the CV space (6). Because the scope of the overall procedure is to construct a model to describe the thermodynamic and kinetic properties of the system, it is important that the microstates be defined in such a way that they satisfy three properties: (*i*) the microstates should densely cover all the configuration space explored in BE-META, including the barrier regions; (*ii*) the distance in CV space between the centers of the nearest-neighbor microstates should not be too large; and (*iii*) the population of each microstate in the BE-META trajectory has to be significant, otherwise its free-energy estimate will be unreliable. A set of microstates that satisfy these properties is defined by dividing the CV space in small hypercubes forming a regular grid. The size of the hypercube is defined by its side in each direction: $ds = (ds_1, ds_2, \ldots, ds_n)$, where $n$ is the number of CVs used in the analysis. This procedure directly determines how far the cluster centers are in CV space. Each frame of the BE-META trajectory is assigned to the hypercube to which it belongs, and the set of frames contained in a hypercube defines a cluster.

The free-energy $F_\alpha$ of each microstate $\alpha$ is estimated by a weighted-histogram analysis (WHAM) approach (17), as described previously (6). In the WHAM approach, the effect of the bias is removed, thus resulting in the free energy of a finite number of microstates that are representative of all the configurations explored by the system. The free energy of a microstate $\alpha$ is given as

$$F_\alpha = -T\log\frac{\sum_i n^i_\alpha}{\sum_j e^{\frac{1}{T}\left(f^j - V^j_\alpha\right)}}, \qquad \textbf{[S11]}$$

where $n^i_\alpha$ is the number of times the microstate $\alpha$ is observed in the trajectory $i$ and $V^i_\alpha$ is the bias potential acting on microstate $\alpha$ in the trajectory $i$. $V^i_\alpha$ is estimated as the time average of the history-dependent potential acting on the trajectory $i$, evaluated in $s_\alpha$, the center of microstate $\alpha$:

$$V^i_\alpha = \overline{V^i_G}(s_\alpha) = \frac{1}{t_{sim} - t_{eq}}\int_{t_{eq}}^{t_{sim}} dt' V^i_G(s_\alpha, t'), \qquad \textbf{[S12]}$$

where $t_{sim}$ is the total simulation time and $t_{eq}$ is the time after which the bias potentials become stable. The normalization constants $f^j$ appearing in Eq. **S11** are determined self-consistently as in the standard WHAM method (6). Corrections taking into account the variation of the bias over different structures assigned to the same cluster $\alpha$ also were described previously (6).

An important issue is how many and which CVs should be used in the procedure. It is not necessary to use all the CVs that have been explicitly biased in one replica, as some of these CVs might prove to be a posteriori less relevant for the process or to be strongly correlated with other variables. The variables used for the analysis must provide an accurate and effective description of the system. An accurate description entails a set of microstates in which each member contains consistently similar structures and thus has very similar free energy compared with $k_BT$. If the variables are too few, a microstate will contain structures that are very different from one another. On the other hand, performing the analysis in a very high-dimensional CV space will lead to poor statistics. A graphical user interface for VMD (18) is available that helps one make this choice by easily visualizing the structures assigned to each microstate for different choices of CVs (19).

1. Laio A, Parrinello M (2002) Escaping free-energy minima. *Proc Natl Acad Sci USA* 99(20):12562–12566.
2. Laio A, Gervasio F (2008) Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep Prog Phys* 71:126601.
3. Bussi G, Laio A, Parrinello M (2006) Equilibrium free energies from nonequilibrium metadynamics. *Phys Rev Lett* 96(9):090601.
4. Piana S, Laio A (2007) A bias-exchange approach to protein folding. *J Phys Chem B* 111(17):4553–4559.
5. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314:141–151.
6. Marinelli F, Pietrucci F, Laio A, Piana S (2009) A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLOS Comput Biol* 5(8):e1000452.
7. Pietrucci F, Laio A (2009) A collective variable for the efficient exploration of protein beta-sheet structures: Application to SH3 and GB1. *J Chem Theory Comput* 5:2197–2201.
8. Bonomi M, et al. (2009) PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput Phys Commun* 180:1961–1972.
9. Hess B, Kutzner C, Van Der Spoel D, Lindahl E (2008) Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4:435–447.
10. Martoňák R, Laio A, Parrinello M (2003) Predicting crystal structures: The Parrinello-Rahman method revisited. Beta-sheet structures: Application to SH3 and GB1. *Phys Rev Lett* 90:75503.
11. Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J Am Chem Soc* 131(39):13894–13895.
12. Pople JA (1958) Molecular orbital theory of aromatic ring currents. *Mol Physiol* 1:175–180.
13. Robustelli P, Kohlhoff K, Cavalli A, Vendruscolo M (2010) Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure* 18(8): 923–933.
14. Camilloni C, Robustelli P, De Simone A, Cavalli A, Vendruscolo M (2012) Characterization of the conformational equilibrium between the two major substates of RNase A using NMR chemical shifts. *J Am Chem Soc* 134(9):3968–3971.
15. Ulrich EL, et al. (2008) BioMagResBank. *Nucleic Acids Res* 36(Database issue): D402–D408.
16. Shen Y, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105(12):4685–4690.
17. Kumar S, Bouzida D, Swendsen R, Kollman PA, Rosenberg J (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J Comput Chem* 13:1011–1021.
18. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graph* 14(1):33–38, 27–28.
19. Biarnés X, Pietrucci F, Marinelli F, Laio A (2012) METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. *Comput Phys Commun* 183:203–211.

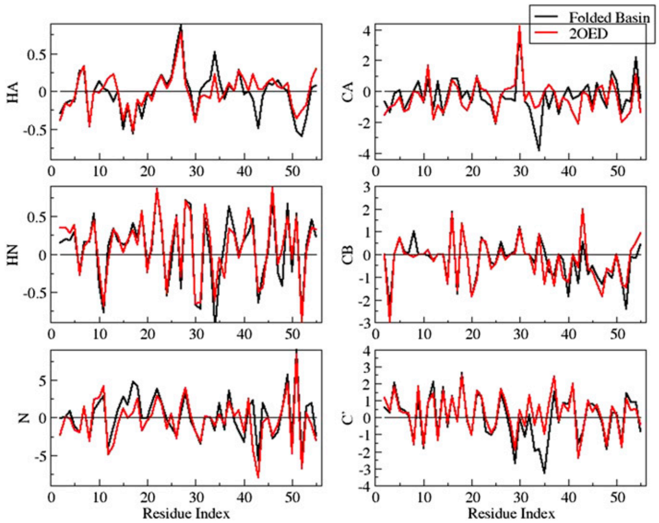

**Fig. S1.** Difference between the chemical shifts calculated with the CamShift method (11) and the corresponding experimental values (16) for the structures in the free-energy minimum (black line) and for the experimental structure (PDB ID code 2OED, red line) reference. The values on the *y* axis are in parts per million.

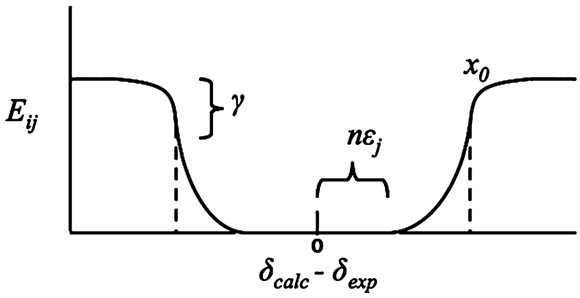

**Fig. S2.** Graphical representation of the functional form of $E_{ij}$ used to calculate the CamShift CV (adapted from ref. 13).

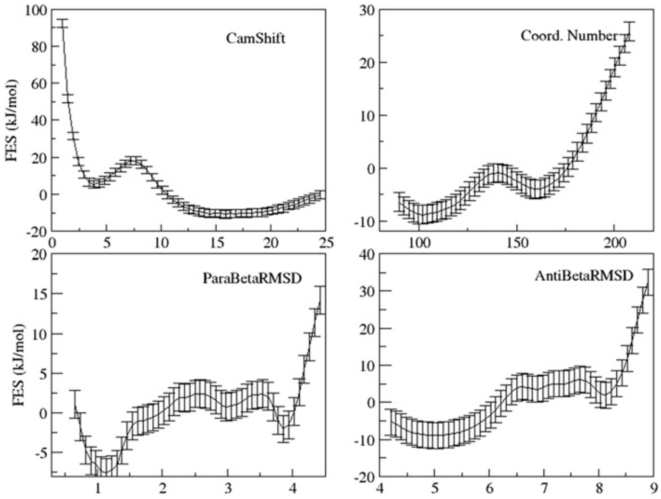

**Fig. S3.** Free-energy surfaces (FES) as function of different collective variables for GB3.

**Table S1. Comparison between the maximum force introduced by a single Gaussian $\left(\frac{w}{\sigma_s}\right)$ and the average of the derivative of the free energy with respect to the specific CV $\left(\overline{\left|\frac{\partial F}{\partial s}\right|}\right)$**

| CV ($s$) | $w/\sigma_s$, kJ/mol | $\overline{\lvert\partial F/\partial s\rvert}$, kJ/mol |
|---|---|---|
| CamShift ($\sigma_s$ = 1) | 0.3 | 3.4 |
| CamShift ($\sigma_s$ = 0.5) | 0.6 | 3.4 |
| Coordination Number ($\sigma_s$ = 10) | 0.03 | 0.35 |
| ParaBetaRMSD ($\sigma_s$ = 0.1) | 3.0 | 13.7 |
| AntiBetaRMSD ($\sigma_s$ = 0.2) | 6.0 | 13.0 |

$w$ = 0.3 kJ/mol.