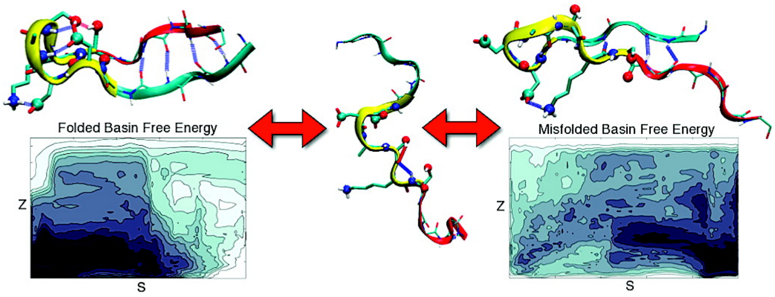
Article

The Unfolded Ensemble and Folding Mechanism of the C-Terminal GB1 #-Hairpin

Massimiliano Bonomi, Davide Branduardi, Francesco L. Gervasio, and Michele Parrinello

J. Am. Chem. Soc., **2008**, 130 (42), 13938-13944 • DOI: 10.1021/ja803652f • Publication Date (Web): 24 September 2008

Downloaded from http://pubs.acs.org on February 8, 2009



Journal of the American Chemical Society is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

JACS ARTICLES

Published on Web 09/24/2008

The Unfolded Ensemble and Folding Mechanism of the C-Terminal GB1 β-Hairpin

Massimiliano Bonomi, Davide Branduardi,* Francesco L. Gervasio,* and Michele Parrinello

Computational Science, Department of Chemistry and Applied Biosciences, ETH Zürich, USI Campus, Via Giuseppe Buffi 13, CH-6900 Lugano, Switzerland

Received May 16, 2008; E-mail: davide.branduardi@phys.chem.ethz.ch; francesco.gervasio@phys.chem.ethz.ch

Abstract: In this work, we shed new light on a much-studied case of β-hairpin folding by means of advanced molecular dynamics simulations. A fully atomistic description of the protein and the solvent molecule is used, together with metadynamics, to accelerate the sampling and estimate free-energy landscapes. This is achieved using the path collective variables approach, which provides an adaptive description of the mechanism under study. We discover that the folding mechanism is a multiscale process where the turn region conformation leads to two different energy pathways that are connected by elongated structures. The former displays a stable 2:4 native-like structure in which an optimal hydrophobic packing and hydrogen bond pattern leads to 8 kcal/mol of stabilization. The latter shows a less-structured 3:5 β-sheet, where hydrogen bonds and hydrophobic packing provide only 2.5 kcal/mol of stability. This perspective is fully consistent with experimental evidence that shows this to be a prototypical two-state folder, while it redefines the nature of the unfolded state.

Introduction

The discovery of the first water-soluble β-hairpin—namely, the C-terminal domain (residues 41−56) of the immunoglobulin binding protein GB1 (GB1P)—paved the way to understanding the factors that govern the β-sheet structure.[1,2] Its well-characterized kinetic and thermodynamic properties, its fast folding times of 6 μs,[3] and its importance as a prototype of a β-hairpin structure have attracted vast interest from both the computational chemist[4−26] and the experimental biochemist community.[1,3,27−30] The folding mechanism and factors contributing to the stabilization of the GB1P are much clearer now. Equilibrium and kinetics studies using temperature jump,[3] calorimetric, and nuclear magnetic resonance (NMR) experiments,[1,27,31,32] supported by calculations and simulations, showed that this peptide forms the secondary structure cooperatively and that its folding can be depicted as a two-state process.[3,4,33]

Both the turn and interstrand hydrophobic side-chain−side-chain interactions have been shown to contribute to the β-hairpin

[1] Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584–590.
[2] Hughes, R.; Waters, M. *Curr. Opin. Struct. Biol.* **2006**, *16*, 514–524.
[3] Muñoz, V.; Thompson, P.; Hofrichter, J.; Eaton, W. *Nature* **1997**, *390*, 196–199.
[4] Muñoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci., USA* **1998**, *95*, 5872–5879.
[5] Kolinski, A.; Ilkowski, B.; Skolnick, J. *Biophys. J.* **1999**, *77*, 2942–2952.
[6] Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci., USA* **2000**, *97*, 2544–2549.
[7] Dinner, A.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci., USA* **1999**, *96*, 9068–9073.
[8] Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151–169.
[9] Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci., USA* **1999**, *96*, 9062–9067.
[10] Roccatano, D.; Amadei, A.; Di Nola, A.; Berendsen, H. J. C. *Protein Sci.* **1999**, *8*, 2130–2143.
[11] García, A.; Sanbonmatsu, K. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 345–354.
[12] Zhou, R.; Berne, B.; Germain, R. *Proc. Natl. Acad. Sci., USA* **2001**, *98*, 14931–14936.
[13] Zhou, R.; Berne, B. *Proc. Natl. Acad. Sci., USA* **2002**, *99*, 12777–12782.
[14] Bolhuis, P. *Proc. Natl. Acad. Sci., USA* **2003**, *100*, 12129–12134.
[15] Bolhuis, P. *Biophys. J.* **2005**, *88*, 50–61.
[16] Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci., USA* **2005**, *102*, 6801–6806.
[17] Daidone, I.; D'Abramo, M.; Nola, A. D.; Amadei, A. *J. Am. Chem. Soc.* **2005**, *127*, 14825–14832.
[18] Ma, B.; Nussinov, R. *J. Mol. Biol.* **2000**, *296*, 1091–1104.
[19] Lee, J.; Shin, S. *Biophys. J.* **2001**, *81*, 2507–2516.
[20] Ma, B.; Nussinov, R. *Protein Sci.* **2003**, *12*, 1882–1893.
[21] Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci., USA* **2004**, *101*, 14766–14770.
[22] Colombo, G.; DeMori, G. M. S.; Roccatano, D. *Protein Sci.* **2003**, *12*, 538–550.
[23] Yoda, T.; Sugita, Y.; Okamoto, Y. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 846–859.
[24] Wang, H.; Sung, S. S. *Biopolymers* **1999**, *50*, 763–776.
[25] Wei, G.; Mousseau, N.; Derreumaux, P. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 464–474.
[26] Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *121*, 1080–1090.
[27] Honda, S.; Kobayashi, N.; Munekata, E. *J. Mol. Biol.* **2000**, *295*, 269–278.
[28] Du, D.; Zhu, Y.; Huang, C.; Gai, F. *Proc. Natl. Acad. Sci., USA* **2004**, *101*, 15915–15920.
[29] Du, D.; Tucker, M.; Gai, F. *Biochemistry* **2006**, *45*, 2668–2678.
[30] Munoz, V.; Ghirlando, R.; Blanco, F.; Jas, G.; Hofrichter, J.; Eaton, W. *Biochemistry* **2006**, *45*, 7023–7035.
[31] Blanco, F.; Serrano, L. *Eur. J. Biochem.* **1995**, *230*, 634–649.
[32] Wei, Y.; Huyghues-Despointes, B. M. P.; Tsai, J.; Scholtz, J. M. *Proteins* **2007**, *69*, 285–296.
[33] Bussi, G.; Gervasio, F.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435–13441.

10.1021/ja803652f CCC: $40.75

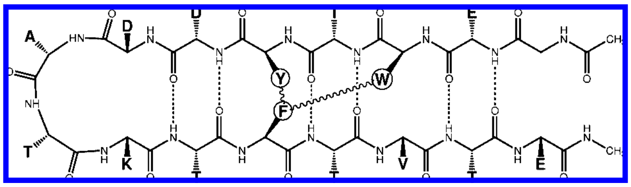


Figure 1. Schematic representation of the β-hairpin structure in the folded conformation.

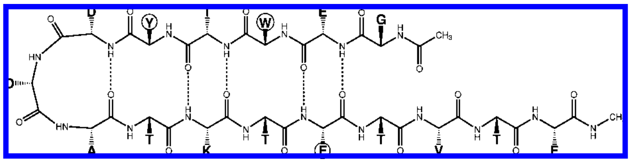


Figure 2. Schematic representation of the β-hairpin structure in the misfolded conformation.

stability.[18,28,29,34,35] The relative contributions of these components and the detailed folding mechanism are still debated.

Much less attention has been paid to the nature of the unfolded ensemble, despite the fact that, in 1995, Frank et al.[36] and Blanco and Serrano[31] already demonstrated the existence of residual structure under denaturing conditions, by studying the whole GB1 and the isolated fragment 41−56 in urea solution. The existence of some form of structure in the denatured state is corroborated by recent findings on other β-hairpins (trpzip and TC5b), which also show residual structure.[29,37,38] These results call into question the naive notion that the unfolded ensemble is composed of fully disordered and stretched structures.

In a previous calculation based on a combination of metadynamics and parallel tempering,[33] we showed that the free-energy landscape can indeed be partitioned into two states: a folded region and an unfolded one. The nature of the latter, however, was never fully elucidated, neither in our studies nor in many other similar theoretical studies. Here, the implicit assumption also was that the unfolded states could be described as highly disordered with a large gyration radius. Here, we address this issue with a set of calculations based on the path method of Branduardi et al.[39] and by re-examining the results of previous simulations.[33] In such a way, we gain a detailed understanding of the nature of the unfolded state and of the folding process as well. This is done based on a fully atomistic description of the protein and the solvating water molecules. Our results can be summarized as follows. A fully stretched configuration is unstable toward the formation of a turn, which is an event that occurs in a rather short time span. The turn can assume two different conformations: a native-like one, which eventually leads to the 2:4 native structure, and a non-native turn, which characterizes the ensemble of the unfolded states among which an ordered 3:5 misfolded structure is particularly stable. Thus, a free-energy landscape emerges in which the two main basins can be identified: a folding basin (Figure 1) and a misfolded one (Figure 2), separated by a transition region of fully stretched configurations.

Simulation Details

The GB1P (PDB code: 1GB1) was solvated in 1559 tip3p water molecules[40] and three Na ions were added to ensure charge neutrality. To preserve the charge distribution of the natural system,[12] we used the non-zwitterionic form of the protein (Ace-GEWTYDDATKTFTVTE-Nme). The OPLSAA[41] force field was used throughout, as implemented in version 2.6 of the NAMD[42] molecular dynamics code. The temperature was enforced using a Langevin thermostat. The free-energy surfaces were determined by metadynamics[43] in its direct formulation.[44] Gaussians with a height of 0.1 kcal/mol were deposited each picosecond. We departed from the usual procedure of keeping the Gaussian width constant and determined the widths from the fluctuations in the values of the collective variables. We checked explicitly that this does not introduce appreciable errors while accelerating considerably sampling convergence. The fluctuations were obtained from averages performed during the interval between one deposition and the other. The simulation length was 180 ns for each metadynamics run. Each free-energy profile was obtained via the multiple walkers technique,[45] with the number of walkers ranging from four to eight.

The Path Method

Path-inspired collective variables were used throughout to study the folding mechanism. We briefly recall the main features of the path collective variables (PCV) approach. We consider a case in which we study a transition between the stable or metastable states A and B.[46] We assume that the transition from A to B can be described by a set of collective variables $\mathbf{S}(\mathbf{R})$ that are generally nonlinear vectorial functions of the microscopic variables $\mathbf{R}$. In our case, they will be the coordinate of a subset of atoms or the elements of the contact map. Contrary to the metadynamics, the dimension of the vector $\mathbf{S}$ is not restricted to a small number but can be arbitrarily large. If the choice of $\mathbf{S}$ is appropriate, we would expect the reactive trajectories to be bundled in a narrow tube around a path, which we write in parametric form as $\mathbf{S}(t)$ for $0 \le t \le 1$ with $\mathbf{S}(0) = \mathbf{S}_A$ and $\mathbf{S}(1) = \mathbf{S}_B$. The minimum free-energy path that connects A to B on the free-energy surface $F(\mathbf{S})$ is of great interest:

$$F(\mathbf{S}) = -\frac{1}{\beta}\ln\langle\delta(\mathbf{S} - \mathbf{S}(\mathbf{R}))\rangle \tag{1}$$

where the average is taken over the Boltzmann distribution. As shown by Ren and Vanden-Eijnden,[47] such a path also has

[34] Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci., USA* **2001**, *98*, 5578–5583.

[35] McCallister, E. L.; Alm, E.; Baker, D. *Nat. Struct. Biol.* **2000**, *7*, 669–673.

[36] Frank, M. K.; Clore, G. M.; Gronenborn, A. M. *Protein Sci.* **1995**, *4*, 2605–2615.

[37] Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153–164.

[38] Mok, K. H.; Kuhn, L. T.; Goez, M.; Day, I. J.; Lin, J. C.; Andersen, N. H.; Hore, P. J. *Nature* **2007**, *447*, 106–109.

[39] Branduardi, D.; Gervasio, F.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 054103.

[40] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

[41] Kaminsky, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

[42] Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[43] Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci., USA* **2002**, *20*, 12562–12566.

[44] Laio, A.; Fortea-Rodriguez, A.; Gervasio, F.; Ceccarelli, M.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109*, 6714–6721.

[45] Raiteri, P.; Laio, A.; Gervasio, F.; Micheletti, C.; Parrinello, M. *J. Phys. Chem. B* **2006**, *110*, 3533–3539.

[46] Bolhuis, P.; Chandler, D.; Dellago, C.; Geissler, P. *Annu. Rev. Phys. Chem.* **2002**, *54*, 20.

[47] Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; Weinan, E. *J. Chem. Phys.* **2005**, *123*, 134109.

dynamical meaning under the appropriate and often-realized conditions. To trace this path, we follow the procedure of Branduardi et al.[39] and introduce the two variables $s(\mathbf{R})$ and $z(\mathbf{R})$:

$$s(\mathbf{R}) = \lim_{\lambda \to \infty} \frac{\int_0^1 t\mathrm{e}^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(t)\|^2}\,\mathrm{d}t}{\int_0^1 \mathrm{e}^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(t)\|^2}\,\mathrm{d}t} \qquad (2)$$

$$z(\mathbf{R}) = \lim_{\lambda \to \infty} -\frac{1}{\lambda} \ln \int_0^1 \mathrm{e}^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(t)\|^2}\,\mathrm{d}t \qquad (3)$$

With ||...||, we indicate the metric that defines the distance between the configurations. For any microscopic configuration $\mathbf{R}$, $s(\mathbf{R})$ and $z(\mathbf{R})$ measure its intercept and distance from the path $\mathbf{S}(t)$, respectively. In practical applications, we describe the path with a discrete number of frames $\mathbf{S}(l)$ (for $l = 1, ..., P$), with $\mathbf{S}(1) = S_A$ and $S(P) = S_B$. The integrals in eqs 2 and 3 are approximated by the finite sums

$$s(\mathbf{R}) = \frac{1}{P-1} \frac{\sum_{l=1}^{P} (l-1)\mathrm{e}^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(l)\|^2}}{\sum_{l=1}^{P} \mathrm{e}^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(l)\|^2}} \qquad (4)$$

and

$$z(\mathbf{R}) = -\frac{1}{\lambda}\ln\left(\sum_{l=1} P\mathrm{e}^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(l)\|^2}\right) \qquad (5)$$

Care must be taken that all $\mathbf{S}(l)$ are equally spaced, relative to the metric used, and that there is sufficient overlap between the clusters defined by the reference frames. A distinguishing feature of this method is its nonlocal character; in fact, when used together with metadynamics, it is able to find transition paths that are rather different from the initial one. This is achieved by exploring the free-energy dependence on $z(\mathbf{R})$, which is the variable that measures the distance from the reference path.

In this work, we consider two different sets of collective variables. The first one, which was already explored in ref 39 for a simple dipeptide, defines $\mathbf{S}(\mathbf{R})$ as a set of the Cartesian coordinates of a subset of atoms. The distance between different configurations ||...||, to be used in eqs 4 and 5, was then measured as the root-mean-square deviation (rmsd) between the two structures after they have been optimally aligned using the Kearsley algorithm.[48] In a second choice, $\mathbf{S}(\mathbf{R})$ is given by the $j > i$ elements of the contact map (CMAP) matrix $\mathbf{C}(\mathbf{R})$, which is defined as

$$\mathbf{C}(\mathbf{R})_{i,j} = \frac{1-\left(\frac{r_{i,j}}{r_0}\right)^6}{1-\left(\frac{r_{i,j}}{r_0}\right)^{10}} \qquad (6)$$

where $r_{i,j}$ is the distance between the ith and jth C_α atoms of the protein backbone and, following Vendruscolo,[49] the distance r_0 is taken to be $r_0 = 8.5$ Å.

Our definition of the contact map is formally different from that commonly used in the literature,[49] which is discrete, and where r_0 is intended as a sharp cutoff but it is identical in its spirit. Here, in fact, we must define a contact in terms of a differentiable function with a continuous derivative, because metadynamics requires smooth forces in the full range of distances.

The square distance, $\|...\|^2$, between a generic state R and a point along the path described by the CMAP $\mathbf{S}_\mathrm{C}(l)$ is, in this case, measured as

$$\|\mathbf{S}_\mathrm{C}(\mathbf{R}) - \mathbf{S}_\mathrm{C}(l)\|^2 = \sum_{j>i} (\mathbf{C}_{i,j}(\mathbf{R}) - \mathbf{C}_{i,j}(l))^2 \qquad (7)$$

where the nearest neighbors are excluded from the sum.

To distinguish between the two cases that have been discussed, we shall use the suffix R and C from this point forward to represent the variable s and z to indicate that we are using the rmsd and CMAP metrics, respectively.

In ref 39, we have shown that the free-energy surface $F(s, z)$ offers precious informations on the possible reaction paths. Furthermore, an efficient procedure was proposed to improve the initial path until it lies on a minimum free-energy line that connects A with B on the free-energy surface. This amounts to minimizing the tension functional,

$$T = \int F(s, 0)\,\mathrm{d}s \qquad (8)$$

relative to the reference path $\mathbf{S}(t)$ through the variables s and z. (See Supporting Information).

To obtain the starting reference path $\mathbf{S}(t)$, the 10 most-populated clusters lying along the L-shaped minimum obtained by parallel tempering[33] were extracted. This seemed to be the most natural choice, given the conventional wisdom. We anticipate that this was not the optimal one. Remarkably, however, the path method itself was able to guide us toward a more-correct choice.

These clusters were ordered and interpolated, to obtain P equally spaced configurations to be used in eqs 4 and 5. The chosen values of λ were 2.81 Å^{-2} for the rmsd metric and 4.98 for the CMAP metrics. Unless otherwise specified, we have used $P = 10$, which ensures that, when measured with the rmsd criterion, successive configurations along the path are not further away than 1 Å, which is slightly smaller than the usual 1.5 Å cutoff distance to distinguish between different clusters. A representative sketch of the set of 10 configurations used to define $\mathbf{S}_\mathrm{C}^0(l)$ is shown schematically in Figure 3. Note that, in the final unfolded configuration, the intramolecular hydrogen bonds are all broken except for two bonds in the turn region from residue 46 to residue 51, with this being the last structured highly populated cluster that lies beyond the transition state.

Before setting up a metadynamics run, it is expedient to determine the maximal range of collective variables that it is useful to explore. To this effect, we run an 8-ns long simulation at 500 K, and this will help us set up bounds on the maximum allowed values for the variables z_C. These are enforced by repulsive harmonic walls placed at $z_C = 8$.

Results and Discussion

In the preliminary high-temperature run, a complete unfolding was obtained, reaching a final state with an rmsd value from the native fold of >10 Å. In several runs of 5 ns at $T = 300$ K, starting from this configuration, the spontaneous formation of the turn was observed. Sometimes, the turn was properly folded, whereas, at other times, it assumed a different hydrogen bond pattern. We took these observations as an indication that the turn arrangement could play a major role. A fast collapse of stretched structures has also been reported in an extensive set

[48] Kearsley, S. K. *Acta Crystallogr., Sect. A* **1989**, *45*, 208–210.

[49] Vendruscolo, M.; Najmanovich, R.; Domany, E. *Phys. Rev. Lett.* **1999**, *82*, 656–659.

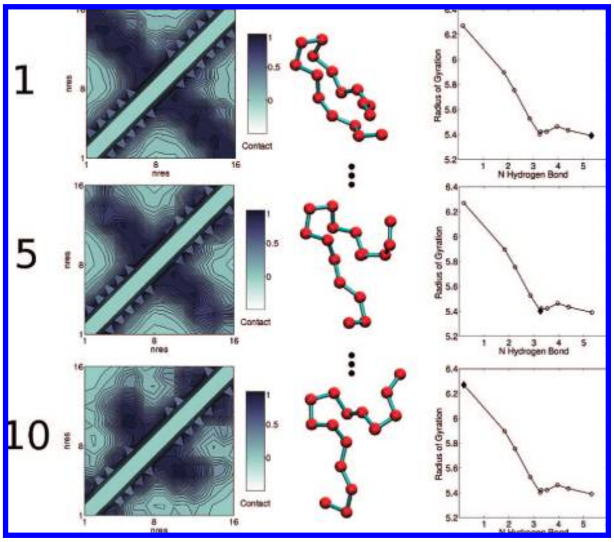


Figure 3. Representative snapshot of the CMAP path. Left panels: examples of the contact maps used. Center panels: representative sketches of the corresponding C_α configurations. Right panels: plots of the number of native hydrogen bonds versus the radius of gyration of the hydrophobic residues for the reference frames.

of calculations[37] on a β-sheet-forming peptide similar to the one simulated here. We decided to simulate the GB1P with PCV in the CMAP space, because this is expected to sample the configurational space of the unfolded regime more effectively. As described in the Methods section, the starting reference path $\mathbf{S}_C^0(l)$ was built on the free-energy surface obtained in ref 33 and optimized. After 10 optimization steps, a satisfactory convergence was achieved, as measured by

$$\sum_m \left| \sum_l \frac{\partial F(l,0)}{\partial \mathbf{S}(m)} \right| \tag{9}$$

A metadynamics run conducted with these variables showed several folding and refolding processes. It is remarkable to observe that, in these successful refolding events, the rmsd value from the NMR folded structure was as low as 1.8 Å, including the heavy atoms of the side chains, which were not included in the definition of s_C and z_C.

However, after ~150 ns, the system was pushed to a region of configuration space from which refolding of the protein was no longer observed.

The nature of the event that triggers the irreversible behavior becomes clear if we monitor the rmsd from the native structure of the heavy atoms from residues 46 to 51 that form the turn region of the protein, as shown in the bottom panel of Figure 4. The abrupt jump seen at ~150 ns corresponds to the disruption of the native fold of the turn and the formation of a different hydrogen bond pattern, leading to a misfolded turn, in agreement with our preliminary unbiased molecular dynamics runs. From this point forward, the system explores the high z_C region of the configuration space, where occasionally visits an incorrectly 3:5 pattern folded state.[21,25,26] This is another indication that the way the turn is folded plays an important role as suggested, but not proven, by many authors.[28-30] This prompted us to study in detail this transition.

To zoom into this transition, it is convenient to redefine the reference path. Thus, we take as collective variables only the heavy atoms involved in the formation of the turn-stabilizing hydrogen bond pattern (see Figure 5). The local character of the turn breaking and reforming event suggests using the more shortsighted rmsd metric, rather than the more global CMAP approach. The reference path is constructed by extracting equally spaced frames from a metadynamics trajectory between 125 and 140 ns (see Figure 4). If we apply, to this case, the criterion that successive frames should have a rmsd value of <1 Å, we should use the values $P = 20$ and $\lambda = 7.5$ Å^{-2}.

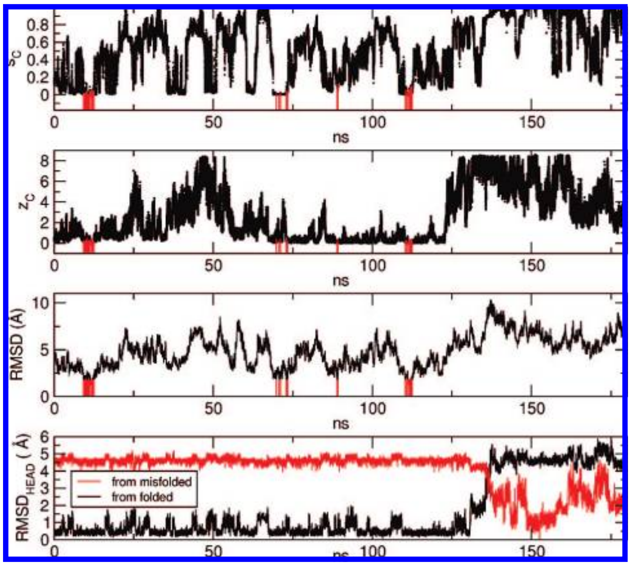


Figure 4. Plot of the timeline during metadynamics in 180 ns: the top panel shows the variable s_C and the second panel shows the variable z_C. The folding events are shown in red, where the rmsd value, relative to all the heavy atoms of the protein, is <1.8 Å (shown in the third panel). The corresponding points are shown as red lines in the upper panels. The bottom panel shows the irreversible event with the rmsd, with respect to all heavy atoms of the native structure in the 46−51 region. The red line corresponds to the rmsd from the misfolded conformation in the 46−51 region.

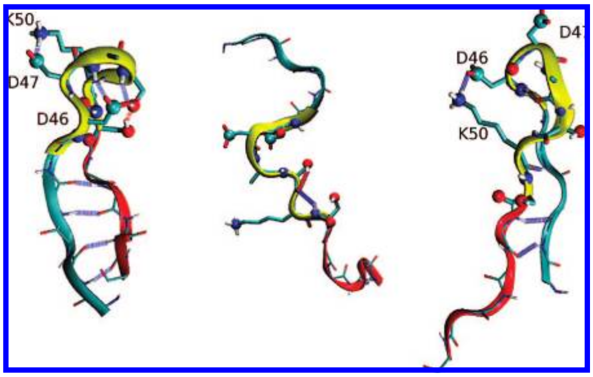


Figure 5. Representation of the folded−unfolded−misfolded transition. Bigger spheres represent atoms involved in the definition of the turn s_R variable.

We then proceed to optimize this path. Although convergence required a considerable number of optimization steps (~200), the qualitative nature of the transition did not change. The z_R does not introduce any appreciable features; therefore, we show the free-energy profile $F(s_R) = -kT \ln \int e^{-\beta F(s_R,\, z_R)}\, dz_R$ that is associated with the optimized free-energy path thus obtained (see Figure 6). Full convergence of the free energy for $s_R < 0.35$ and $z_R > 0.8$ proved to be difficult, because of the role of relatively slow orthogonal degrees of freedom, such as hydrophobic side-chain interactions and backbone hydrogen bonds not pertaining to the turn region. Only the region $0.35 < s_R < 0.8$ could be properly sampled, leading to the free energy depicted in Figure 6. It is reassuring in this respect to report

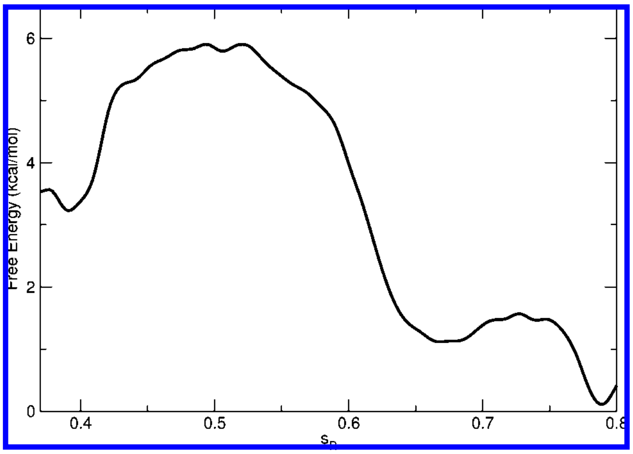


Figure 6. Free-energy profile as a function of the path in the turn region.

that, in this range, several unfolding and refolding processes were observed.

We can roughly distinguish three regimes: (i) the initial one, which has the lowest free energy and in which the turn is well-formed ($s_R \leq 0.1$), (ii) a transition region ($0.1 < s_R \leq 0.6$), in which the protein fully unfolds; and (iii) a final one ($s_R > 0.6$) in which the turn is incorrectly folded.

To understand the nature of this conversion mechanism, an analysis of different contributions to the free energy is estimated (see Supporting Information for calculation details). It turns out that the potential energy increases substantially, because of the loss of intraprotein hydrogen bonds. This effect is not counterbalanced by the increasing entropic term of the highly disordered extended structured, thus producing the free-energy barrier of the transition.

Analysis of the Parallel Tempering Results

These results are reinforced by an analysis of a previous unbiased parallel tempering run.[33] Using the configurations thus obtained, we calculated the free energy, with respect to the variables s_R and z_R, which describe the turn conformation. In Figure 7, these are compared to the now-standard description where the free energy is mapped out as a function of the number of hydrogen bonds and the gyration radius of the hydrophobic residues. We first focus on the $T = 300$ K free energies (panels 1 and 4). It is seen that s_R and z_R are coherent with the results described in the previous section. All but a handful of states can be classified based on the turn fold, either as “folded-turn” or “misfolded-turn”-like.

The parallel tempering data at the higher temperatures also becomes more instructive if plotted as a function of s_R and z_R. The unstructured configurations at $s_R \approx 0.5$ become more populated. However, even at the highest temperature, the folded and misfolded basin are still clearly evident.

Three Different Regions

Until now, we have determined three different regions in the free-energy surface (FES) of the β-hairpin, distinguished by the conformation of the turn. To characterize them, we have run three PCV simulations in CMAP space: one keeping the turn close to the native conformation ($F^F(s_C, z_C)$), one keeping it close to the misfolded conformation ($F^M(s_C, z_C)$), and the last keeping the turn unfolded ($F^U(s_C, z_C)$). This is done by imposing three different constraints ($s_R \leq 0.1$, $0.4 \leq s_R \leq 0.6$, and $s_R \geq 0.6$, which amount to selecting states in which the turn is correctly formed, totally disrupted, or misfolded, respectively.

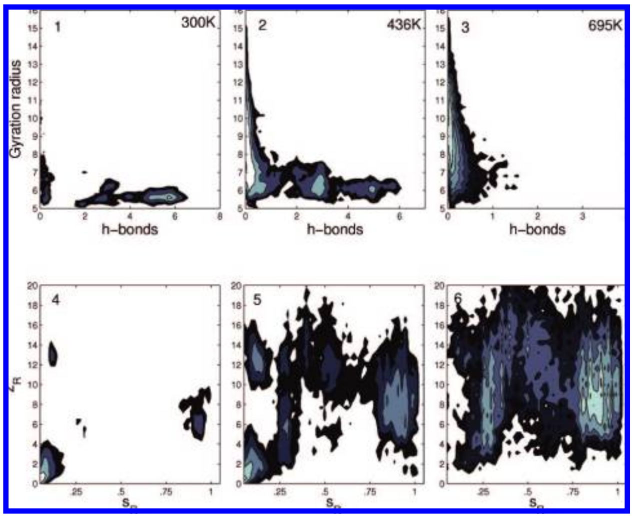


Figure 7. Upper row: free energy profiles, as a function of the number of hydrogen bonds (horizontal axis) and the gyration radius of the hydrophobic residues (vertical axis), as obtained from a 8.4-ns parallel tempering run.[33] Lower row: corresponding free-energy profiles as a function of s_R and z_R of the turn region. The contours are spaced at intervals of 1 kcal/mol. The three rows represent three different temperatures, as indicated in the upper panels.

Let us first consider $F^F(s_C, z_C)$ (see Figure 8). In this case, we use the optimized reference path illustrated in Figure 3. It can be seen that $F^F(s_C, z_C)$ clearly shows that the folded structure is the free-energy minimum. A clear funnel structure is apparent, despite the distortion induced by the mapping into the s_C and z_C variable. In $F^F(s_C, z_C)$, three different patches can be identified. Patch A, which includes the folded structure, has the largest number of native hydrogen bonds and the smallest gyration radius. In patch B, the number of hydrophobic contacts is reduced more than that of the native hydrogen bonds, while the protein itself retains a compact structure. Finally, in patch C, the protein is on its way to the transition state and the number of both native hydrogen bonds and hydrophobic contacts is smaller, while the gyration radius is larger.

The rich structure in $F^F(s_C, z_C)$ may be contrasted with the featureless flatland of $F^U(s_C, z_C)$, where once the turn is unfolded, no preference for specific structures is to be seen (see Figure 9).

We now turn to the study of $F^M(s_C, z_C)$. Clearly, we cannot use the reference path of Figure 3 to describe this basin. Actually, if we attempt this, the system has a clear tendency to explore large-z regions, indicating that the path is far from optimal.

On the other hand, during the turn fold-driven metadynamics path, we found a perfect 3:5 β-hairpin structure shown in Figure 5 (right panel), so it was natural to investigate its stability. Thus, we performed a 10-ns run during which this “perfectly” misfolded structure remained stable. Therefore, it is a legitimate metastable state to be considered.

Thus, the new path was built from the 3:5 β-hairpin to a structure analogous to the final structure of $F^F(s_C, z_C)$. This final structure lies close to the transition state of Figure 6.

With this choice, the metadynamics runs in a smooth, reversible way and does not show any anomalies. The corresponding free-energy surface again exhibits a funnel character (shown in Figure 10). In this case, however, only the non-native

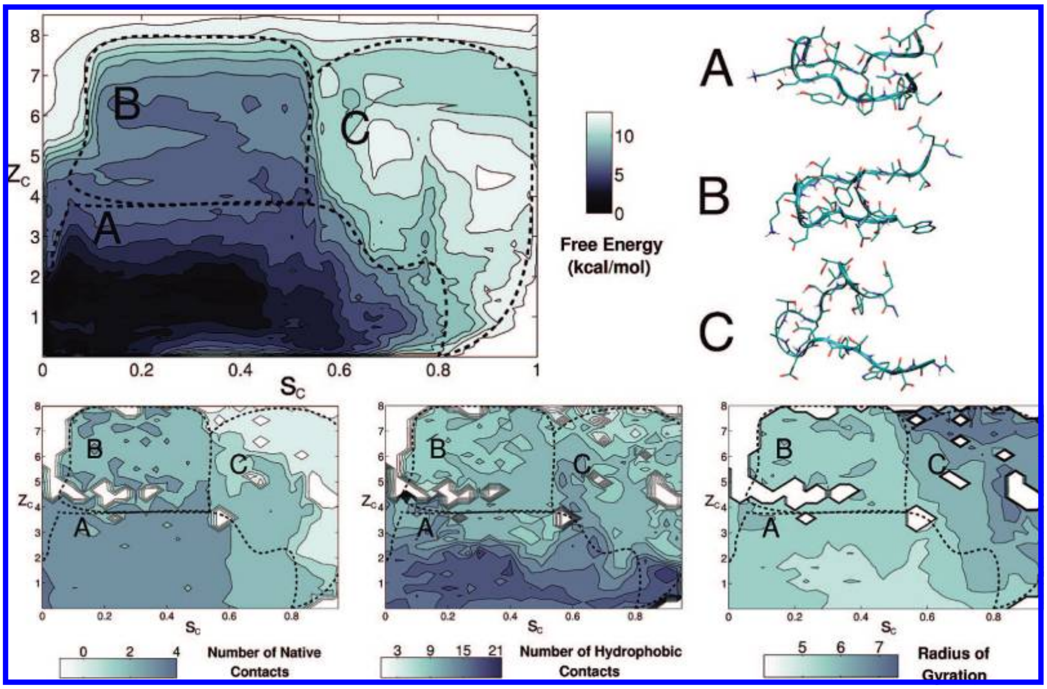


Figure 8. Description of $F^F(s_C, z_C)$, showing the free-energy landscape as a function of s_C and z_C and representative plots of the native hydrogen bonds, hydrophobic contacts, and the gyration radius of the hydrophobic residues for the properly formed turn region (s_R < 0.1). Contours are drawn every 1 kcal/mol. In ribbons: three typical structures encountered in basins A, B, and C.

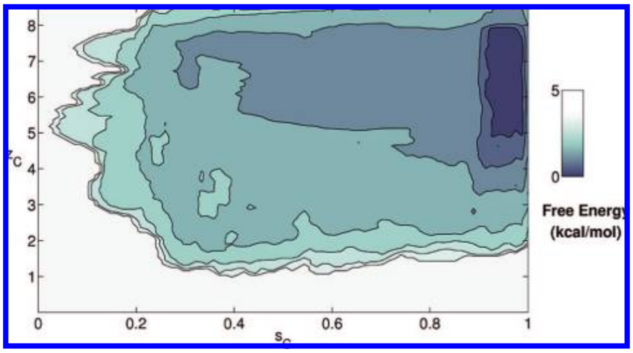


Figure 9. Description of $F^U(s_C, z_C)$, showing the free-energy surface in the saddle region (0.4 < s_R < 0.6). Contours are drawn every 0.5 kcal/mol.

hydrogen bond stabilizes the 3:5 β-hairpin, because the misfolded turn hampers the tight packing of the hydrophobic residues. This leads to a shallower funnel, (at its deepest, only 2.5 kcal/mol high).

Compiling all this information, we find that, in the free energy landscape of this small protein, one can identify two main funnels: one that leads to the native folded structure and another that, instead, leads to a 3:5 misfolded one.

We now ask whether a direct path exists that brings the folded structure to the totally misfolded one without passing through extended structures. Indeed, this "reptation-like" movement has been proposed in ref 25, based on coarse-grained simulations. Therefore, we have constructed such a hypothetic path by joining together a series of highly populated structures that bind the folded structure to the incorrectly unfolded one. The path is optimized and the final free-energy surface is drawn in Figure 11. It can be seen that there is a free-energy path, but this does not run close to the $z_C = 0$ axis. The low free-energy path is far from the reference path and involves the complete denaturation of the protein, passing from elongated states in which even the turn region is unfolded. This confirms our previous results and strengthens the picture of a free-energy space, which can be partitioned into two regions: one in which the turn is well-formed, and another in which the turn is misfolded. To go from one region to the other, the turn must be unfolded.

Conclusions

We have studied, in depth, the room-temperature unfolded ensemble and the folding mechanism of the C-terminal domain (residues 41−56) of the immunoglobulin binding protein GB1. To this end, we used state-of-the-art methods such as our recently developed path-like collective variables and parallel-tempering simulations, producing several trajectories more than 180 ns long. This procedure makes use of an adaptive set of collective variables that, in perspective, could describe the folding mechanism of other small peptides. Our results are in agreement with several lines of experimental evidence in showing that GBP1 is a two-state folder. However, we have shown that the "unfolded" state can be characterized more by the structure of the turn than by its compactness. Indeed, we find strong evidence of two main energy funnels that, depending on the conformation of the turn, lead either to the folded structure or to a misfolded structure.

In this small peptide, once the complete ”denatured” state, which is intended to be a coil-like structure, is characterized, its probability turns out to be low, with respect to the folded and misfolded basins. Similar conclusions were drawn by van Gunsteren et al.[51] for several small polypetides.

The crucial role of the turn is consistent with experiments on the entire protein and on the isolated fragment. On the entire protein, mutagenetic experiments by the Baker group showed that a key residue of the turn is essential to the stability.[35] On the fragment, many studies based on different experimental

[50] Kobayashi, N.; Honda, S.; Yoshii, H.; Munekata, E. *Biochemistry* **2000**, *39*, 6564–6571.

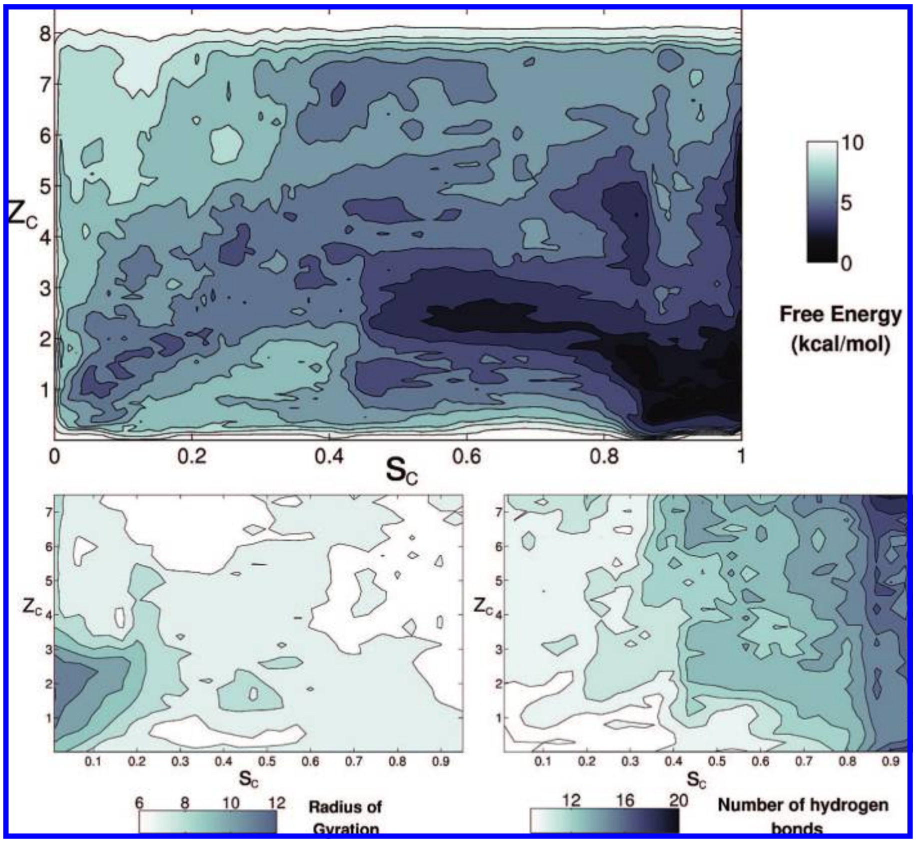


Figure 10. Description of $F^M(s_C, z_C)$, showing the free-energy landscape as a function of s_C and z_C and representative plots of hydrogen bonds and gyration radius for the misfolded turn region (s_R > 0.6). Contours are drawn every 1 kcal/mol.

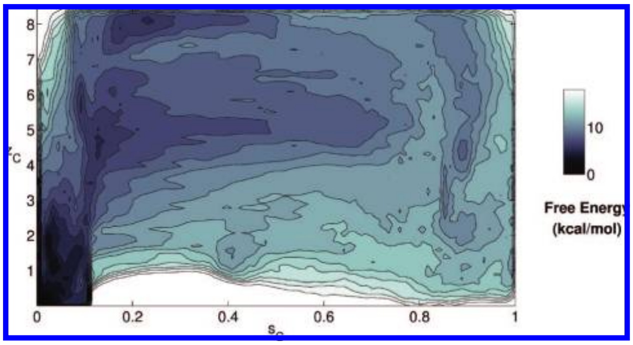


Figure 11. Description of the reptation mechanism: the points at $z_C \simeq 0$ are not explored, confirming that the sliding mechanism is not feasible for the wild type. Contours are drawn every 1 kcal/mol.

techniques such as nuclear magnetic resonance (NMR),[1,31] laser temperature jump,[3] differential scanning calorimetry (DSC),[50] and φ-value analysis[29] have also come to the conclusion that mutating the turn residues leads to a change in the kinetics of the folding process. Experimental evidence supporting the existence of a well-defined structure in the misfolded region is provided in the work of Serrano and co-workers. They used trifluoroethanol as their solvent, which is known to stabilize the secondary structure. These authors, in fact, reported, in this solvent, two NOE peaks, which were assigned to the formation of a Gly41−Val54 bond and Glu42−Val54 bond. Such an interaction is weak in the perfect fold but it is fully compatible with our misfolded structure. It would be illuminating to reexamine all these data on the basis of the present picture.

Acknowledgment. The Swiss National Supercomputing Center (CSCS) is kindly acknowledged for providing computational time under the ALPS scheme. We would also thank Giovanni Bussi for many stimulating discussions.

Supporting Information Available: Details on the frames optimization procedure and an estimate of the entropy contribution for the folded to misfolded transition is given (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

JA803652F

(51) van Gunsteren, W. F.; Bürgi, R.; Peter, C.; Daura, X. *Angew. Chem., Int. Ed.* **2001**, *40*, 351–355.