Advillin Folding Takes Place on a Hypersurface of Small Dimensionality

Stefano Piana
Nanochemistry Research Institute, Curtin University of Technology, Perth, Western Australia

Alessandro Laio
SISSA/ISAS Via Beirut 2-4 Trieste, Italy

(Received 5 February 2008; published 10 November 2008)

All-atom explicit-solvent molecular dynamics simulations have been used to investigate the topological structure of the space explored during folding by the *c*-terminal fragment of the Advillin headpiece, a 36 amino-acid protein. A fractal dimension analysis shows that the hypersurface explored during the folding process has an approximate dimensionality of only three. It is shown that this low dimensionality persists well above the unfolding temperature and is not present in simple coarse-grained models.

DOI: 10.1103/PhysRevLett.101.208101 PACS numbers: 87.15.Cc, 87.15.ap, 87.15.hm

In 1969 Levinthal proposed, in his famous paradox, that protein folding should take an exceedingly long time to occur because of the astronomically large number of conformations that have to be explored during the folding process [1]. Several theories have been put forward during the years to solve this paradox. It has been proposed that the folding process does not proceed as a random walk in the space of conformations but is guided towards the native state by a funnel-shaped free energy landscape [2,3] in which transitions towards nativelike structures are marginally but systematically favored. It has also been proposed that the effective number of configurations that the system has to explore before finding the native state is greatly reduced by the constraints imposed by the intrinsic features of polypeptide chains that, for instance, bend with a fixed radius of curvature and form hydrogen-bonds with fixed patterns [4,5]. The importance of constraints in determining the dynamics of proteins is confirmed by some recent simulation results. It has been shown that two generalized reaction coordinates are sufficient to capture the most important features of the folding landscape of SH3 and CV-N proteins described with a coarse-grained model [6]. Moreover, it was recently demonstrated by extensive explicit-solvent simulations that the dynamics of poly-Ala takes place on manifolds of reduced dimensionality [7].

Here extensive atomistic molecular dynamics simulations have been used to further investigate the topological features of the configuration space that is explored by a protein during the folding process. The analysis is performed on the Advillin *c*-terminal headpiece (Advillin), a 36 amino-acid (AA) protein that folds to form a three helix bundle [8]. The small size of the system allowed performing extensive simulations using an accurate but computationally expensive description that takes into account the solvent molecules explicitly [9]. The simulation was performed using bias exchange metadynamics [10] (BE) and run for 640 ns on 8 replicas (details can be found in Ref. [11]). BE, by enforcing an extensive exploration of the configuration space, allows predicting the folded state of small proteins. A cluster analysis [10] performed on the BE trajectory allows identifying hundreds of structures that differ in secondary content, solvent accessible surface area, number of internal hydrogen bonds, radius of gyration, etc. For each of these structures the free energy was estimated from the BE results [11]. The most stable structure has a root mean square deviation of 2.5 Å with respect to the experimental structure [8], showing that the simulation predicts the correct fold [11].

The trajectory obtained from this simulation was used to calculate the fractal dimension of the space of the conformations explored during the folding process. To this aim, the trajectory was projected in the space defined by the seven collective variables that are used in the BE simulation (other choices of variables will be considered below). These variables are the number of backbone hydrogen bonds, salt bridges and hydrophobic contacts, the number of α/β residues and the correlation between the backbone dihedral angles [10,11]. The last two variables are estimated separately for the first and the last 18 residues of the protein. Each of the seven variables is defined as a continuous function of the coordinates [10,11] changes significantly during the dynamics and does not show any obvious correlation with the others: the plot of the trajectory as a function of any pair of variables looks invariably like a dense two-dimensional region, indicating that the dimensionality of the configuration space explored during folding is at least two. Among the several possible measures of the fractal dimension of a series of observables, the correlation dimension [12] was chosen here, as it was found to provide the best convergence with respect to the number of data and was the easiest to compute for high-dimensional spaces. Test calculations performed with the box counting method gave similar results, but convergence was not nearly as good. The correlation dimension is a measure of how the number of neighbors of a point in-

creases with distance [12]. Consider a hypersurface of dimension d embedded in a higher dimensional space. The number of points within distance r from a given data point scales as r^d where d is, by definition, the correlation dimension. d can be estimated from the slope of the natural logarithm of the number of neighbors $N(r)$ versus the natural logarithm of r. Such a plot is reported in Fig. 1 (green points). Two different correlation exponents can be spotted, with a crossover at $\log(r) \sim -2.7$. For $\log(r) < -2.7$, corresponding to small conformational changes, the correlation dimension of the space is ~ 6, close to the maximum possible value of 7. However, for larger conformational changes the correlation dimension of the space becomes as low as 3.02. The small size of Advillin (36 AA) limits the size of the space where large conformational changes can be observed. Still, the behavior of $\log(N(r))$ is compatible with a dimensionality of ~ 3 for more than 1 order of magnitude in the variation of r. Qualitatively, the space of conformation of this protein can be described as local regions of high-dimensional space embedded in a low-dimensional superstructure. This means that, on average, each configuration can evolve in only three linearly independent directions, and the number of pathways that the system can follow is much reduced. This is consistent with the results obtained studying the large-scale motion of globular proteins by normal mode analysis [13] and with a recent analysis performed on poly-Ala [7], SH3, and CV-N proteins [6].

The robustness of this result was tested in several different manners. First, the estimate was repeated eliminating half of the data points, one out of every two. The new plot (Fig. 1, blue) is indistinguishable from the calculation based on the full set (Fig. 1, green), showing that the result is well converged with respect to the number of data. Second, the analysis was repeated on a completely independent BE simulation, initiated from a different starting structure. This second simulation also correctly predicts that the most stable structure of the system is the native fold. The correlation dimension that is obtained (red points in Fig. 1) is essentially identical to the one obtained in the first simulation. Finally, the correlation dimension was computed on a 4 μs normal (unbiased) molecular dynamics trajectory of Villin at 340 K kindly provided by Eastwood and Shaw. This trajectory starts from an extended state, explores a smaller portion of configuration space than the BE trajectory, but finally visits the folded state. The intrinsic dimension of the space sampled by this simulation is 2.8, indicating that the low dimensionality is not an artifact generated by the BE algorithm. This also suggests that the effective dimensionality of the folding space is determined by the gross features of the Hamiltonian as a complete exploration of the configuration space does not seem to be essential to estimate its value. The correlation dimension was also computed separating the trajectory in two parts, one including foldedlike structures at a RMSD from the folded state below 5 Å, the other including all the rest. Once again, the correlation dimension in the two parts is indistinguishable, indicating that the small value is not a specific feature of the folded minimum.

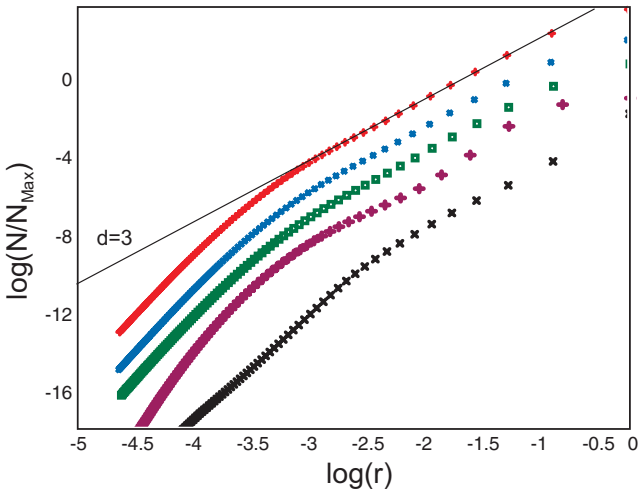

FIG. 1 (color). Plot of the logarithm of the number of neighbors N as a function of the logarithm of the distance r. Since r has to be defined in a space including variables with different units of measure, before computing the distance each variable s is divided by $s_{\max} - s_{\min}$, where $s_{\max}$ and $s_{\min}$ are the maximum and the minimum value of s explored during the dynamics. Data are reported for simulation 1 (green), simulation 2 (red), half of the data of simulation 2 (blue), simulation 2 in a 30-dimensional representation based on the number of $C\alpha$ contacts of residues 4–33 (black) and simulation 2 in a 15-dimensional space based on the $C\alpha$-$C\alpha$ distances (purple). The black line corresponds to a correlation dimension of 3 and is plotted as a guide for the eyes. For the sake of clarity, the curves obtained in the different simulations are displaced by an arbitrary constant in y direction, otherwise some of them would be indistinguishable.

The results discussed so far were obtained computing the correlation dimension on a trajectory embedded in a 7-dimensional space defined by the variables used to perform the BE simulation. One might wonder if the dimensionality of ~ 3 that is found depends on this choice. It is obvious that adding a coordinate that is irrelevant for the folding (e.g., the rotation of a methyl group or the position of a solvent molecule) has the effect of increasing the observed dimensionality by one. However, the rotation of a methyl group or the position of a water molecule are *fast* variables that are explored very efficiently also in a short time. It can be checked that as long as the correlation dimension is computed in a space including only variables that describe global structural rearrangements its value is approximately 3 irrespectively of the set of variables that is chosen. For example, the trajectory was mapped in the 30-dimensional space of the number of contacts n_i that each $C\alpha$ makes with all the other $C\alpha$-s, for each residue starting from 4 up to 33. n_i is estimated as a continuous function of the positions r_i of the $C\alpha$-s as

$$n_i = \sum_{j=4}^{33} \frac{1-(r_{ij}/r_0)^{10}}{1-(r_{ij}/r_0)^{12}}$$

where $r_0 = 6.5$ Å. The correlation dimension calculated in this high-dimensional space is still 3.2 (Fig. 1, black), very close to the result obtained in the 7-dimensional space. Using the number of $C\alpha$ contacts as collective variables changes the position of the crossover between the high-dimensional and low-dimensional region. However, this is expected, as the position of the crossover is related to the definition of distance, which has a different meaning in each space.

In this respect it is instructive to compute the correlation dimension in a space of variables including only distances, as this allows providing a physical interpretation of the crossover observed in Fig. 1. At this scope, the BE trajectory was mapped in a 15 dimensional space defined by the set of distances r_{ij} between pairs of $C\alpha$-s, with $i = 5$, 10, 15, 20, 25, 30, and $j = i + 5, \ldots, 30$. The distance in this space between the two structures at time t and t' is defined as $r = \sqrt{\sum_{ij}(r_{ij}(t) - r_{ij}(t'))^2}$ and is closely related to the standard RMSD between the two structures. The correlation dimension computed in this set of variables displays a more complex behavior, but is still approximately 3 over a wide range of distances (purple crosses in Fig. 1). The crossover between the lower and higher dimensionality is located at $\log(r/r_{\max}) \sim -2.5$, corresponding to $r \sim 1.6$ Å. This suggests that local rearrangements can take place, on average, in several independent directions. As soon as the system moves significantly, say of more than 2 Å in this space, it can move only in less independent directions, approximately between 2.5 and 3.5. The crossovers that are observed at larger r are smeared out in the space of the other collective variables used in Fig. 1 and could be related to the relatively small length of the protein that is considered here, 36 AA. The nature of these crossovers should be analyzed further computing the correlation dimension on folding trajectories of larger proteins.

The temperature dependence of the correlation dimension was estimated by running 30 ns of parallel tempering (PT) simulation on 32 replicas with temperatures ranging between 298 and 480 K [14,15]. To speed up convergence, the central structures of the most populated clusters found in the BE trajectory [10,11] were chosen as initial configurations. As shown in Fig. 2, between 298 and 320 K the correlation dimension increases from ~2 to ~3 and then remains constant up to 380 K, where it starts growing again. Even above the unfolding temperature large portions of the configuration space cannot be accessed and the dimensionality remains rather small. Larger dimensionalities were observed in test simulations where all nonbonded interactions and the dihedral terms involving the C_α-s were eliminated, suggesting that the low d is a consequence of the energetic and geometric constraints imposed by the polypeptide chain. However, the force field used for all-atom explicit-solvent simulations have been optimized to describe accurately the behavior of proteins at room temperature, and the behavior of the model at higher temperatures might be less realistic.

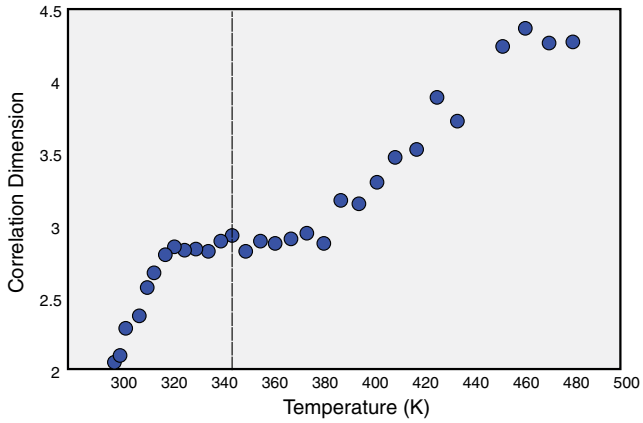

FIG. 2 (color online). Temperature dependence of the correlation dimension of the folding space. The correlation dimension was estimated as in Fig. 1 for each replica of a parallel tempering simulation. The dashed line indicates the melting temperature (342 K).

To investigate further the nature of the low dimensionality the folding of Advillin was also simulated with a Go Hamiltonian [16]. This model, although simple, has been successfully applied for studying the qualitative features of the folding process [17,18]. For the present analysis each residue is represented by a sphere centered on the $C\alpha$ carbon atom. The $C\alpha$-$C\alpha$ interaction is modeled with a 10–12 potential that for $C\alpha$-s closer than 0.65 nm has a minimum of 1 kcal mol^{-1} at the experimental distance and for more distant $C\alpha$-s is repulsive only. Interaction up to the third nearest neighbor were excluded and a force constant of 2 kcal mol^{-1} rad^{-1} and of 0.075 kcal mol^{-1} rad^{-1} was used for the angle terms and for the dihedral terms, respectively. The correlation dimension is calculated in the 15-dimensional space defined by the set of distances between pairs of $C\alpha$ defined above for the atomistic simulations. The simulation was performed at a temperature where the protein is approximately 50% folded. The analysis is performed on a trajectory containing hundreds of folding events. As shown in Fig. 3, the dependence of $\log(N)$ on $\log(r)$ is qualitatively different from the one observed in the all-atom simulations. In the Go model a plateau with a dimensionality of 1.2 is observed for a large range of distances, approximately up to $r = 2$ Å (Fig. 3, X crosses). For larger distances, the dimensionality is significantly larger, of 10 or more. This might lead to conclude that the short-distance correlation dimension of the Go model is very small. Still, further analysis shows that the situation is more complex. Indeed, another major difference between all-atom and Go results

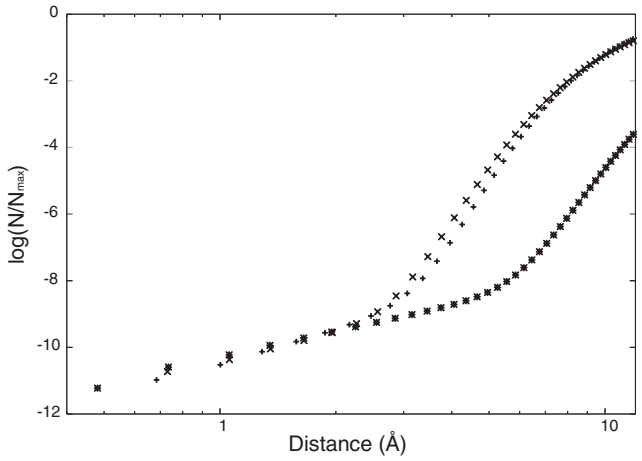

FIG. 3. The logarithm of the number of neighbors N as a function of the distance r in logarithmic scale for a Go model simulations. r is computed in the 15-dimensional space based on the $C\alpha$-$C\alpha$ distances. The correlation dimension was calculated at a temperature where the protein is approximately 50% folded on a trajectory containing hundreds of folding events ($\times$ crosses); same as above, but on a five-times shorter trajectory (pluses); for a purely repulsive self-avoiding polymer (stars). The curve computed with the same set of variables for the all-atom simulation is reported in Fig. 1, purple.

is that the latter converge slowly with simulation times. When the dimensionality analysis is performed on a five-times shorter trajectory, still including several folding events, the low dimensionality region extends toward larger r (Fig. 3, plus). Moreover, if the attractive potential in the Go Hamiltonian is turned off, the one-dimensional region extends towards even larger r (Fig. 3, stars). This corresponds to the correlation dimension curve of a self-avoiding polymer. Also in this case, extending the simulation time has the effect of slowly moving the crossover between the low and the high-dimensional regions towards smaller r. A plausible explanation of this behavior is that the low dimensionality region observed at small distances is a consequence of the high dimensionality at larger r, that would disappear for infinite simulation time. The high dimensionality observed at large distances in the Go model makes it rather unlikely that the system explores several times a configuration that is not the folded state, extremely unlikely if the attractive part of the potential is zero and all the ''unfolded'' configurations are explored, on average, only once. In other words, trajectory recrossings in the Go model simulations are rare and the hypersurface explored during the dynamics in the neighborhood of each configuration coincides with the trajectory itself, whose correlation dimension is, by definition, one. Such a behavior is not observed in all-atom simulations, in which trajectory recrossings are common and only a few folding event are sufficient to converge the results. This suggests that the available space of conformations is in this case intrinsically smaller, and all the relevant conformations are explored a number of times that allow evaluating reliably the correlation dimension at all values of r. It is concluded that the correlation dimension of 3 derives from some nontrivial features of polypeptide chains similar to the ones observed for poly-Ala [7]. These are not captured by a the Go Hamiltonian, that is designed to describe the gross features of the folding process and not to describe the atomistic details of the system.

S. P. acknowledges financial support from the Australian Research Council. Computer time has been provided by APAC and iVEC. The authors are grateful to M. Eastwood and D. E. Shaw for providing the trajectory of their 4 μs simulation of Villin at 340 K. We also acknowledge F. Marinelli, F. Pietrucci, C. Micheletti, A. Maritan, A. Trovato, and J. Gale for several precious suggestions.

[1] C. Levinthal, *How to Fold Graciously, Mossbauer Spectroscopy in Biological Systems*, edited by J. DeBrunner and E. Munck (University of Illinois Press, Monticello, IL, 1969).
[2] P. Leopold, M. Montala, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[3] R. Zwanzig, A. Szabo, and B. Bagchi, Proc. Natl. Acad. Sci. U.S.A. **89**, 20 (1992).
[4] N. C. Fitzkee and G. D. J. Rose, J. Mol. Biol. **353**, 873 (2005).
[5] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).
[6] E. Plaku, H. Stamati, C. Clementi, and L. Kavraki, Proteins: Struct. Funct. Bioinf. **67**, 897 (2007).
[7] R. Hegger, A. Altis, P. Nguyen, and G. Stock, Phys. Rev. Lett. **98**, 028102 (2007).
[8] W. Vermeulen, P. Vanhaesebrouck, M. Van Troys, M. Verschueren, F. Fant, M. Goethals, C. Ampe, J. C. Martins, and F. A. Borremans, Protein Sci. **13**, 1276 (2004).
[9] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplack, R. Luo, and T. Lee. J. Comput. Chem. **24**, 1999 (2003).
[10] S. Piana and A. Laio, J. Phys. Chem. B **111**, 4553 (2007).
[11] S. Piana, F. Marinelli, A. Laio, M. Van Troys, D. Bourry, W. Vermeulen, and J. C. Martins. J. Mol. Biol. **375**, 460 (2008).
[12] P. Grassberger and I. Procaccia, Physica (Amsterdam) **9D**, 189 (1983).
[13] D. Ben-Avraham, Phys. Rev. B **47**, 14559 (1993).
[14] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[15] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[16] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[17] D. Baker, Nature (London) **405**, 39 (2000).
[18] G. Settanni, T. X. Hoang, C. Micheletti, and A. Maritan, Biophys. J. **83**, 3533 (2002).