



Networks of interbasin traffic in intrinsically disordered proteins

Belisa R. H. de Aquino , Mateusz Chwastyk, Łukasz Mioduszewski , and Marek Cieplak*
Institute of Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland



(Received 29 July 2019; accepted 20 December 2019; published 3 March 2020)

The equilibrium dynamics of the intrinsically disordered proteins is thought to consist of transitions between many basins in the free energy landscape whereas structured proteins stay in the vicinity of one native basin. We demonstrate this picture explicitly by studying networks defined on the discretized plane: conformational end-to-end distances vs radii of gyration. The bin sizes are defined by time scales that span orders of magnitude. The networks, derived from all-atom and coarse-grained molecular dynamics simulations, are nearly scale invariant. The bin representation also provides insights into the folding process of the structured proteins and identifies regions of hindrance to folding.

DOI: [10.1103/PhysRevResearch.2.013242](https://doi.org/10.1103/PhysRevResearch.2.013242)

The function of most proteins stems from their specific, experimentally determined spatial structures. It has been recognized, however, that about a third of proteins [1], depending on the organism, lack a single preferred conformation that can be identified with the native state [2]. Such proteins are named intrinsically disordered (IDP) [2–8]. They play important functional roles in the cell, such as signaling, cell-cycle regulation, initiation of transcription/translation, and transport through membranes. They are also involved in neurodegenerative diseases. Their lack of structure results from a deficiency in hydrophobic residues (that could build the hydrophobic core on folding) and an overabundance of charged residues [9].

One of the conceptual issues pertaining to the IDPs is the nature of their dynamics. It is expected that it involves transitions between many basins of attraction, whereas that of the structured proteins amounts to hovering over just one basin that is associated with the native state. In other words, the free-energy landscape of the IDPs is rough and includes many comparable valleys that can be mutually accessed thermally at room temperature, whereas that of the structured proteins is smoother and dominated by one valley. This perspective stems from the past efforts to distinguish sequences corresponding to “bad folders” from those that lead to a folding funnel [10–12]. The case in between is the multifunctional systems [13] when, say, two competing funnels with high separation energy barriers are present as revealed by the disconnectivity graphs [14–18].

The question we pose is how to demonstrate the validity of this picture using molecular dynamics and how to make it quantitative. Simulations generate a string of conformational snapshots. Their interpretation requires simplification,

or coarse graining, of the description. One way to do it is to monitor the evolution of the contact map—the list of contacts detected in a snapshot by using some prescription. This approach would be analogous to a contact-based characterization of the structural effects of mutations [19]. This method, however, leads to a huge number of possibilities even for short chains: for $N = 20$ residues, it is $\sum_{i=0}^m \binom{m}{i}$, where $m = 153$ is the maximal number of contacts excluding the $i, i + 2$ ones (i is the sequence location).

A stronger simplification involves describing the system in terms of global descriptors of the chain, such as the end-to-end distance, L , and the radius of gyration, R_g . These two parameters can be measured in SAXS and FRET experiments. The dynamics can then be represented as a motion on the R_g - L plane as illustrated in Fig. 1 for two systems with $N = 20$: the structured tryptophan cage (Trp-cage) (the structure code is PDB: 1L2Y) and the chain NFGPKGFGYGQAGALVHAQ, which is the 175-194 segment of the disordered cysteine and glycine-rich protein 2 denoted by DP00438 [20]. A sufficient probing yields the equilibrium density of occupation of points on this plane which can serve as a qualitative rendering of the free energy landscape: the most frequented points should correspond to the lowest-energy states. However, such landscapes, *per se*, do not characterize the traffic between possible basins.

Here, we introduce a way to represent the equilibrium dynamics of a protein as a network on the R_g - L plane and demonstrate the existence of a qualitative difference between the networks derived for the structured and disordered systems. The network connects the centers of discrete bins of size $\Delta_R \times \Delta_L$ set on the R_g - L plane. The connectivities are determined through the frequencies, f , of the transitions between the bins. f is defined as the number of occurrences of an interbin transition divided by the number of all transitions. The average occupation, Ω , of a bin is the fraction of time that the system spends in the bin. The most important bins are the hubs of the network.

The crucial aspect of this representation is that the construction of the network depends on the characteristic time scale, Δt , of the description. As the system evolves in time

*mc@ifpan.edu.pl

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

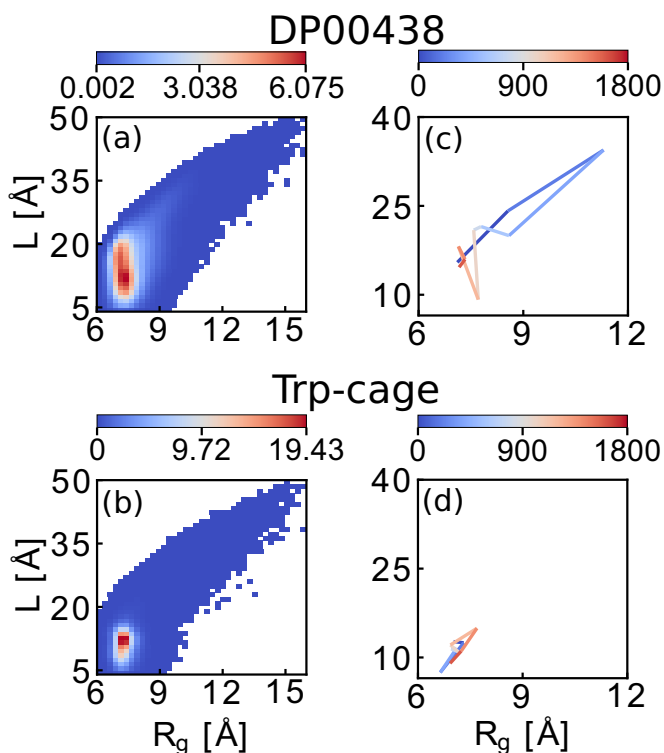


FIG. 1. Panels (a) and (b): accumulated conformational sampling on the R_g - L plane for the systems indicated. The data points were obtained from 100 different trajectories that last for 1 000 000 τ each. The color bars represent the population percentage. The off-diagonal scatter of the data points indicates that the R_g and L are independent variables, as they should be for the description presented in the paper. Panels (c) and (d): the corresponding initial fragments of single trajectories. The cutoff time is 1800 τ . The states are collected every $\Delta t/\tau = 200$. The color bars represent the simulation time interval (in τ).

by Δt , R_g changes by ΔR_g and L by ΔL . These changes are Gaussian distributed with zero mean (Fig. 2). Sampling of Ω is done every Δt . We take the dispersions of these distributions, σ_R and σ_L , as the characteristic sizes of the structural changes that take place in time Δt . The bin sizes are then taken to be as $\Delta R = 2\sigma_R$ and $\Delta L = 2\sigma_L$. The factor of 2 merely reduces the noise. Importantly, the spatial scales are related to the temporal scales. This feature is similar to the free-energy estimates of small molecules being effectively dependent on the time scale of the observation [21].

We illustrate our method by considering the two systems with $N = 20$, Trp cage and DP00438, and two systems with N of about 140. The latter are the structured lysozyme (PDB:2LYZ; $N = 129$) and the disordered [22,23] α -synuclein ($N = 140$), the protein that is associated with Parkinson's disease [24,25]. These proteins are studied by the recently proposed α -C based coarse-grained (CG) model [26] (that is more protein specific than that used in Ref. [27]) and by all-atom (AA) NAMD-based simulations [28] with the implicit solvent and the CHARMM36m [29] force field designed for the IDPs. The CG model is defined in terms of dynamically defined contacts that arise depending on the distance between the residues and on the directions of

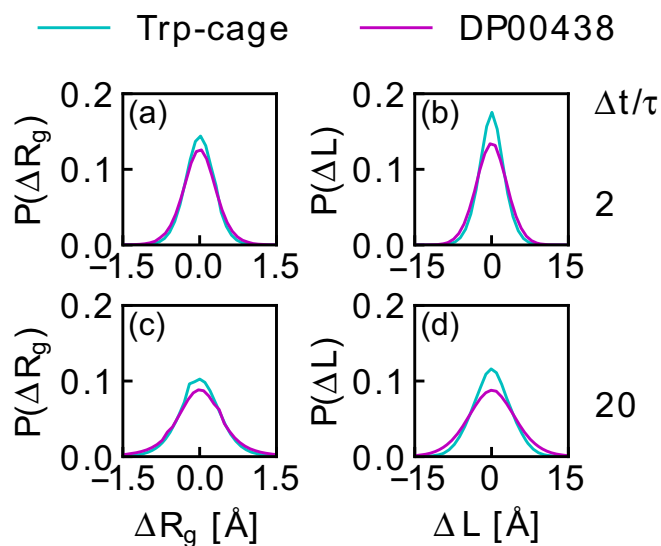


FIG. 2. Probability distributions of ΔR_g and ΔL for Trp-cage and DP00438 for two different timescales. Δt is 2τ in the top panels and 20τ in the bottom panels. The distributions were determined from 100 trajectories lasting for 1 000 000 τ .

the normal and binormal vectors associated with the backbone. There is one change relative to the description provided in Ref. [26]: attractive electrostatic interactions are now treated as other sidechain-sidechain contacts (if the proper conditions are met) instead of being described by the modified Debye-Huckel potential. This correction enhances the agreement with experimental data for charged proteins.

The CG model [26] is defined in terms of ε —the depth of the Lennard-Jones potential well associated with the contact. The room temperature situation corresponds to $k_B T/\varepsilon \approx 0.35$, where k_B is the Boltzmann constant. Most of the calculations were done at this T and the results are based on 100 runs that are 1 000 000 τ long. The backbone stiffness (the bond and dihedral angle terms) are accounted for by statistical potentials. When dealing with the structured proteins, we keep this form of the backbone stiffness but the contacts are preassigned: they are selected based on the existence of overlaps between effective spheres associated with the heavy atoms in the native state [30–32]. The AA simulations were done at the room temperature (298 K) and were obtained in five 30 ns runs. The starting conformations in the CG simulations were self-avoiding random walks. In the AA simulations, the starting conformations were generated through random bending of the backbone by using the Pymol software [33].

Figure 3 shows the networks for Δt of 10 and 20 ps. (All network figures were done by PyGraphviz, pygraphviz.github.io, a Python interface to the Graphviz graph visualization software [34].) The corresponding values of ($\sigma_R/\text{Å}$, $\sigma_L/\text{Å}$) are (0.36, 2.94) and (0.45, 3.46) for the Trp-cage and (0.66, 3.84) and (0.83, 4.83) for DP00438. In the CG case (Fig. 4), the networks are for $\Delta t = 2$ and 20τ , where τ is of order 1 ns. The corresponding values of ($\sigma_R/\text{Å}$, $\sigma_L/\text{Å}$) are (0.29, 2.50) and (0.43, 3.75) for the Trp-cage and (0.43, 3.74) and (0.56, 5.05) for DP00438. The AA length

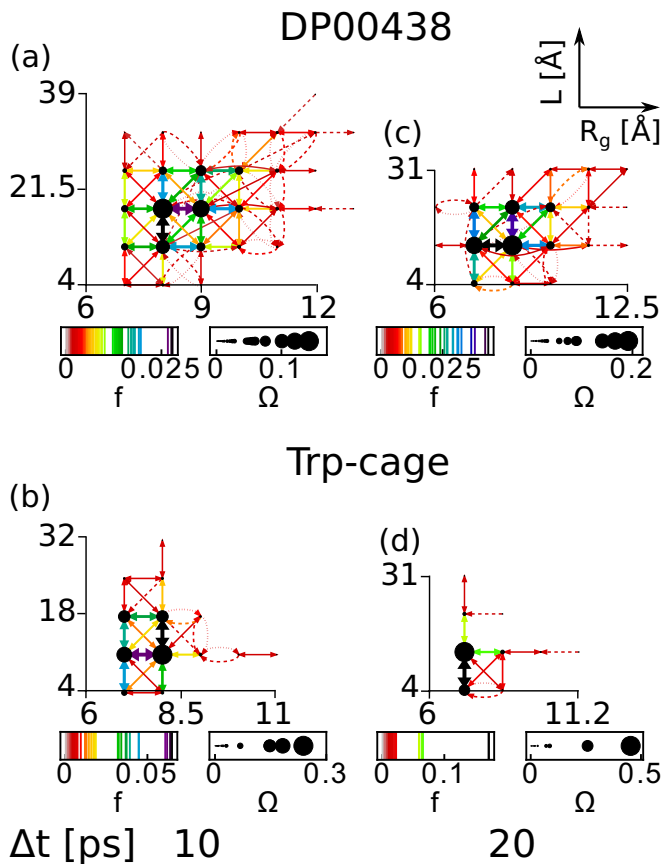


FIG. 3. Network diagrams for DP00438 (top) and Trp-cage (bottom) for $\Delta t = 10$ and 20 ps on the left and right, respectively. The results were obtained through the AA simulations. The values of f and Ω are indicated in the stripes at the bottom of each panel. The color convention used for the network links is shown in the stripes. The black circles indicate the occupational probabilities of the bins. Most of the interbin transitions have symmetric frequencies. They are represented by solid lines with the arrows on both ends. The remaining transitions are indicated by single-arrow lines: the more frequent transition corresponds to the dashed line and the weaker to the dotted line. We show the transitions only with $f > 0.05f_{\max}$, where f_{\max} is the maximal value observed. For Trp-cage f_{\max} is 0.078 and 0.20 for Δt of 10 and 20 ps, respectively. For DP00438 f_{\max} is 0.033 and 0.039, respectively.

scales are mostly somewhat larger than the CG ones, which reflects weaker averaging. We demonstrate the continuity of the behavior across the CG and AA time scales considered.

We study the longer proteins by using the CG model at $\Delta t/\tau = 20$ and we get $(\sigma_R/\text{\AA}, \sigma_L/\text{\AA})$ of (0.21, 1.80) for lysozyme and (0.82, 7.29) for α -synuclein. In order to simplify making comparisons of the same or nearly the same sizes, we average the corresponding dispersions for any given time scale.

Figures 3 and 4 show the networks for the $N = 20$ systems for time scales ranging from 10 ps to $\sim 20\,000$ ps (the results for still larger Δt are not shown as the data are more noisy). We observe that, for a given protein, the look of the network is similar across the time scales. This approximate self-similarity reflects the fact that our description pertains to

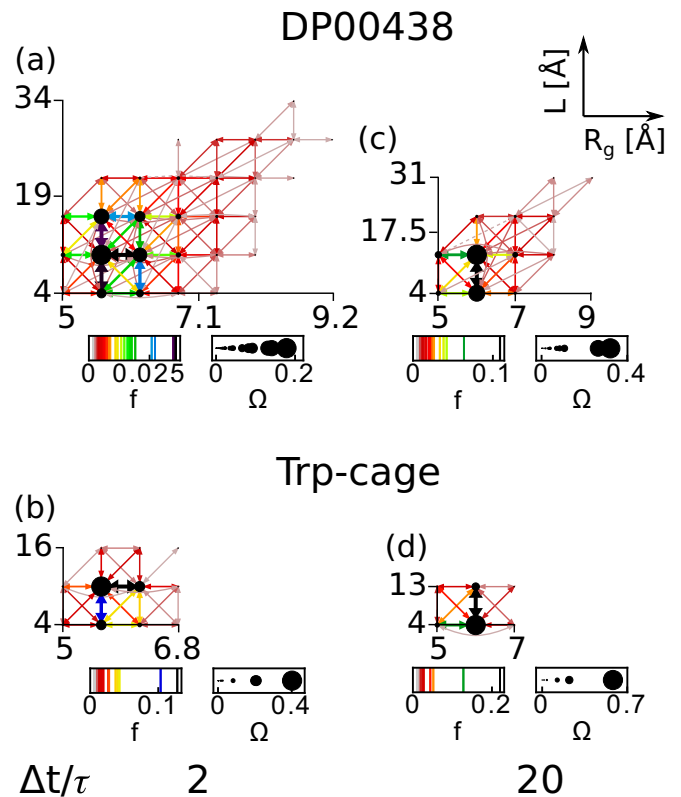


FIG. 4. Similar to Fig. 3 but for $\Delta t = 2$ and 20τ (τ is of order 1 ns) as obtained through the CG simulations. We show the transitions only with $f > 0.01f_{\max}$. For Trp-cage f_{\max} is 0.13 and 0.22 for Δt equal to 2 and 20τ , respectively. For DP00438 f_{\max} is 0.035 and 0.11, respectively.

equilibrium. At the same time, there is a significant qualitative difference between the disordered and structured proteins. The IDP system shows many competing hubs and multiplicity of links. Trp-cage, on the other hand, leads to the emergence of one strongly dominant native hub. The exception is the shortest time scale (10 ps), where several more hubs emerge. This time scale is likely just too short for the system to “realize” its relevant tendencies. However, even at this time scale, the network for DP00438 is substantially more complicated than for Trp-cage. Figure 5 shows that on heating the system to $0.45 \epsilon/k_B$ the systems reduce and reshuffle the weights of the connectivities.

Figure 6 shows that the qualitative difference in the dynamics of the disordered and structured proteins becomes more pronounced when the proteins’ sequences become about seven times longer. It is seen that the network (at $0.35\epsilon/k_B$) for α -synuclein is significantly more interconnected than for lysozyme and there are of order 20 comparably occupied hubs. The network is also much more complex than for DP00438.

One of the important characteristics of a structured protein is its thermodynamic stability. It can be assessed experimentally by a variety of methods, such as circular dichroism [35] and differential scanning calorimetry (DSC) [36]. Typically, these methods are sensitive to only certain segments of the full structure. Theoretically, the thermodynamic stability can

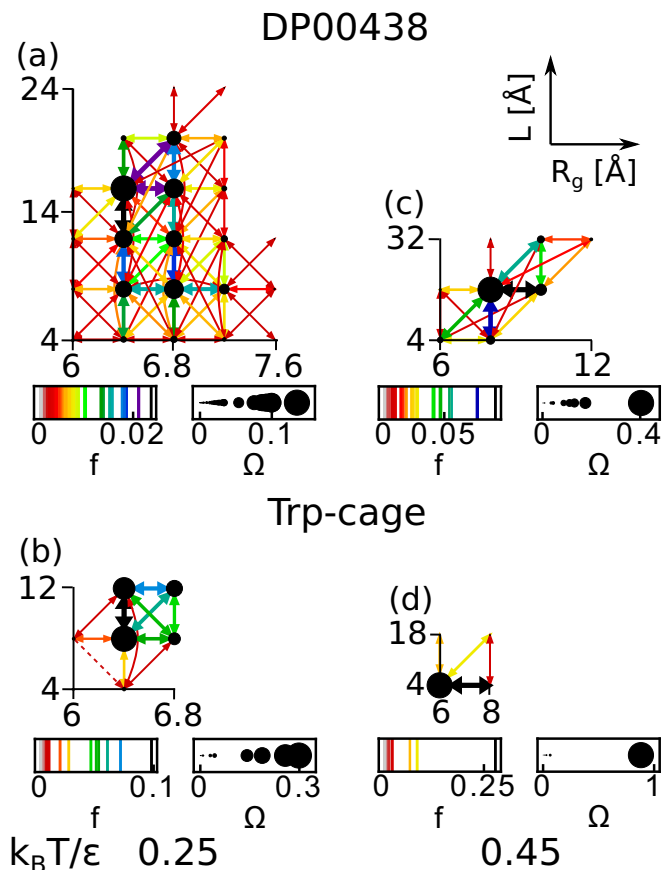


FIG. 5. Behavior of the $N = 20$ systems studied at ε/k_B equal to $0.25 k_B T/\varepsilon$ [panels (a) and (b)] and $0.45 k_B T/\varepsilon$ [panels (c) and (d)]. The bin sizes recalculated for these temperatures. Δt is 20τ . For $k_B T/\varepsilon = 0.25$, f_{\max} and $(\sigma_r/\text{\AA}, \sigma_L/\text{\AA})$ are 0.098 and $(0.19, 1.77)$ for Trp-cage and 0.024 and $(0.19, 2.16)$ for DP00438. For $k_B T/\varepsilon = 0.45$, these are 0.28 and 0.28 and $(0.73, 5.56)$ for Trp-cage and 0.093 and $(1.33, 8.19)$ for DP00438. If one uses the bin sizes corresponding to $0.35 k_B T/\varepsilon$ for the other two temperatures considered here, the networks get modified but preserve similar features (not shown).

be captured by estimating the temperature, T_0 , at which the equilibrium probability of staying in the native state, P_0 , crosses $\frac{1}{2}$. In a contact-based description, P_0 is estimated through the fraction of conformations in which all, or nearly all, native contacts are present simultaneously [37–40]. Assessing the thermal stability in AA models is difficult to achieve in a quantitative way. This is because of the computational cost and the necessity of defining the bounds of the native basin. Thus one typically resorts to nonequilibrium measures such as those obtained through submitting the system to one elevated T and then monitoring the temporal evolution of, for instance, the geometrical distance away from the native structure (RMSD) in unfolding trajectories. This is used in a qualitative way when comparing two systems, e.g., wild type and mutated as in Ref. [41].

For the IDPs, there are additional conceptual problems: can one define a characteristic temperature that provides a measure of the thermal stability despite the absence of the native state? Figure 7 demonstrates that this indeed can be

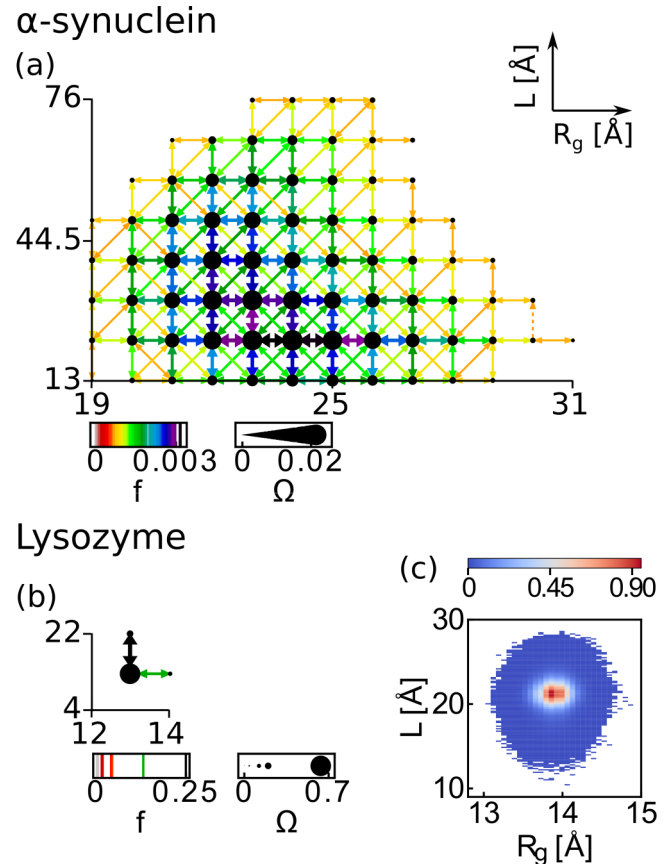


FIG. 6. Similar to Fig. 3 but for α -synuclein (top) and lysozyme (bottom) and only for $\Delta t = 20\tau$ and transitions with $f > 0.2f_{\max}$. The results were obtained by using the CG model. For α -synuclein f_{\max} is 0.0032 and for lysozyme 0.20 . The right bottom panel shows the contour plot for lysozyme.

done by considering the T dependence of average R_g and L . Their values, normalized by the maxima, are denoted here by r_g and l , respectively. Both quantities are seen to display a sigmoidal behavior so we define temperatures T_r and T_l that correspond to the centers of the sigmoids. For a given system, they are fairly close to each other (in units of ε/k_B): 0.40 and 0.40 for DP00438; 0.70 and 0.70 for Trp-cage; 0.40 and 0.40 for α -synuclein; 0.75 and 0.65 for lysozyme. At the same time they differ by ~ 0.3 between the structured and disordered systems and indicate that the structured systems are more stable. In contrast, the specific heat, c_v (Fig. 7), does not differentiate between the structured and disordered proteins of $N = 20$ as the locations of the maxima in c_v are both at 0.35 . However, it does for the two larger proteins: 0.43 and 0.65 for α -synuclein and lysozyme, respectively.

The bins derived in the equilibrium calculations can also be used to characterize nonequilibrium situations such as folding of structured proteins from extended conformations. In our representation, folding is considered to be achieved when the system reaches the native hub for the first time. Figure 8 shows one folding trajectory at $0.35 \varepsilon/k_B$ on the R_g - L plane for lysozyme. The folding time for this trajectory is 3291τ . The

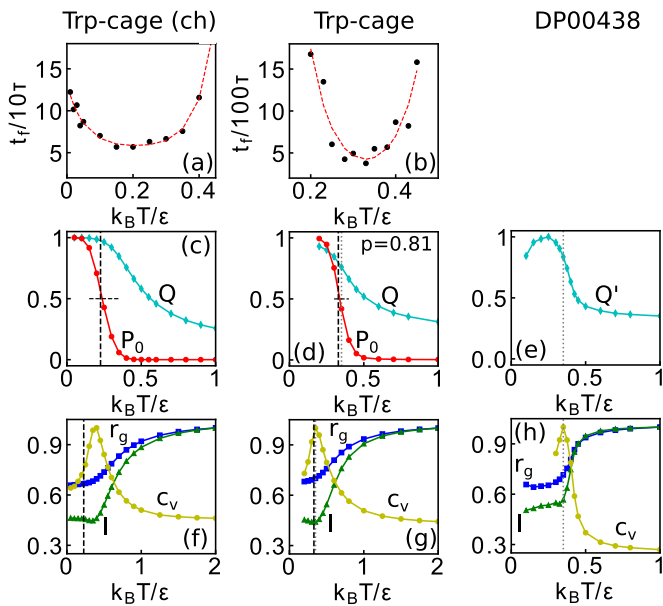


FIG. 7. Temperature dependence of the kinetic and equilibrium parameters calculated for Trp-cage (the first and second column) and DP00438 (the third column). The second and third columns use the statistical potentials in the description of the backbone stiffness. The first column corresponds to the backbone stiffness described by the chirality potential [30,31,40] that favors the specific native conformation. The top panels show the folding times, i.e., the median first passage time needed to establish all native contacts. Notice that the statistical potentials yield folding times that are an order of magnitude longer than with the chirality potential. The middle panels show P_0 , Q , and Q' . Q is the average number of the native contacts (for the structured case) and Q' is the average number of contacts that are present (for the disordered case). The bottom panels show r_g , l , and c_v . All equilibrium quantities are normalized by the maximal values obtained so they do not exceed 1. The vertical dashed lines correspond to the temperature of the folding optimality. This temperature is model dependent. The horizontal dashed lines indicate the value of $P_0 = 0.5$. P_0 is the equilibrium probability of staying in the native state. It is determined by averaging the number of the snapshots in which the native contacts are present simultaneously with a substantial probability [40]. In the model with the chirality, this probability, p , is 1. In the model with the statistical potentials, $p = 0.81$ [40]. The circles correspond to the centers of the corresponding sigmoidal curves.

bins correspond to Δt of 10 and 100τ . The representation of the trajectory depends on Δt and the trajectory itself depends on the starting conformation, indicating multiplicity of the pathways. The 10τ trajectory has two regions of looping movements indicating the existence of metastable traps: R_g of 15–20 Å and 25–32 Å. On increasing the time scale, the upper metastabilities get resolved and the lower region shrinks to 15–17 Å. A further increase in Δt is expected to eliminate the looping entanglements entirely. Thus our method allows one to determine regions corresponding to the folding bottlenecks and time scales needed to overcome them. On averaging over 100 trajectories we get a flowlike pattern (not shown) that is akin to the probability flow studied exactly in a lattice model [39]. For the IDPs, one can study the kinetics of reaching specific hubs.

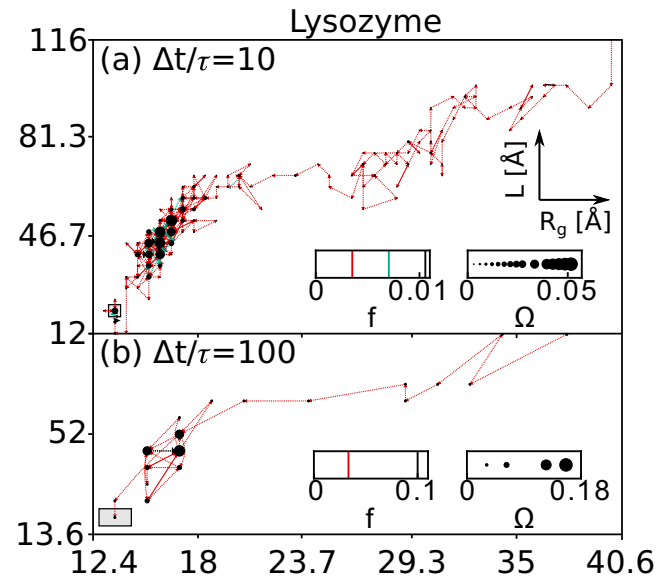


FIG. 8. R_g - L representation of the folding process for lysozyme. The upper/lower panel is for $\Delta t = 10/100\tau$. The color codes for f between the panels are different. For the upper panel, $\Delta R = 0.6$ Å and $\Delta L = 4$ Å. The most occupied bin is centered at (16.6, 52) Å. For the lower panel, $\Delta R = 1.6$ Å, $\Delta L = 6.4$ Å and the largest hub is at (16.2, 45.6) Å. In both cases, the native bin (the gray rectangle) is at (13,20) Å. The values of ($\sigma_R/\text{Å}$, $\sigma_L/\text{Å}$) are (0.33, 2.09) and (0.80, 3.20) for 10 and 100τ , respectively.

Our method to build networks of hubs and links is universal and intuitive. It enhances the notion that the density of points on the R_g - L plane may serve as an effective landscape. Without involving rare-event techniques [42], the resulting free energy—obtained by the Boltzmann inversion—is an upper-bound estimate. Our method allows for a clear elucidation of the differences between the structured and disordered proteins in equilibrium. There are, however, other possible approaches to construct kinetic transition networks [43–47], as reviewed in Ref. [13] for structured proteins. In particular, one can adopt procedures involving a selection of more abstract collective variables that are inferred from eigenvalue decomposition of the raw molecular dynamics data as represented by “feature” vectors. The relevant collective variables can be chosen by maximizing either their variance or their autocorrelation [48,49]. The former are associated with the principal component analysis [50] and the latter with the time-lagged independent component analysis [51–53]. These procedures result in an identification of relevant metastable states and construction of the Markov-state models [54,55] that involve the transitions between the states.

We have applied the TICA approach to our data on α -synuclein and lysozyme and determined that the dynamics of lysozyme are slower: the longest collective time scales differed by at least a factor of 4. This difference appears to indicate that the free-energy landscape for α -synuclein is shallower and better interconnected compared to lysozyme. The landscape for lysozyme is characterized by several deep kinetic traps—one of them corresponds to the mirror image

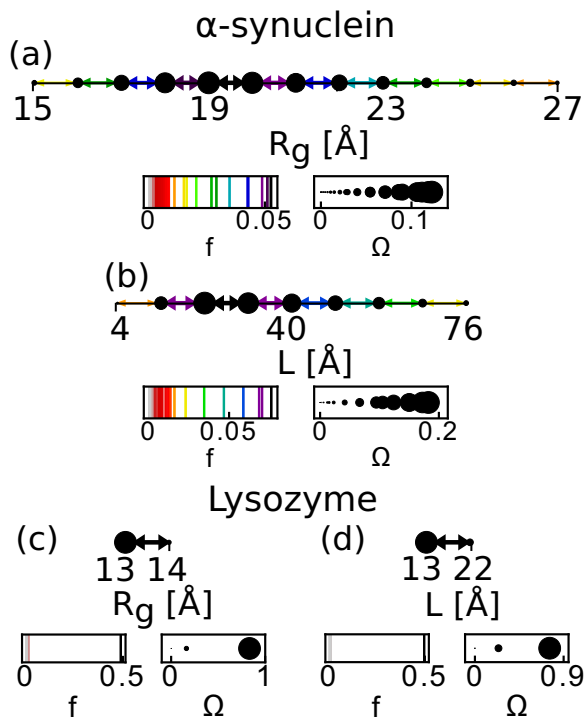


FIG. 9. One-dimensional equilibrium network for α -synuclein (the top panels) and lysozyme (the bottom panels) with $\Delta t = 20\tau$. The traffic is indicated either between the bins along the R_g axis or the L axis. The values of f_{\max} for the R_g -based plots are 0.08 and 0.49 for α -synuclein and lysozyme, respectively. For the L -based plots these are 0.08 and 0.50, respectively.

of the native conformation (the description based on R_g and L cannot distinguish between mirror images).

Other than that, the approach was not found to be very revealing. It should be noted that the collective variables used in the Markov-state models (MSM) depend on the protein which makes comparisons between the proteins indirect. In addition, our bins are of a fixed size and the other descriptors involve variable distances. Our method is helpful in bringing out the role of disorder and in providing a coarse-grained method to represent folding in structured proteins. However, our method is not meant to be used for an identification of a most probable folding pathway since R_g and L are unlikely to correlate with such a pathway strongly. Unlike the disconnectivity-graph approach, it does not operate with a selection of isolated conformations but with ensembles of states associated with the bins. Dealing with the ensembles of states makes it similar to the MSM, although the principles involved are distinct.

A one-dimensional version of our approach (only R_g or only L) is also possible (Fig. 9). Even though our method was defined for a protein chain, it should be generalizable to other systems with complicated free-energy landscapes, such as glasses or protein aggregates.

We appreciate comments from J. R. Banavar. This research has received support from the National Science Centre (NCN), Poland, under Grant No. 2018/31/B/NZ1/00047. This project is a part of a STSM Grant from COST Action CA17139.

- [1] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, *J. Mol. Biol.* **337**, 635 (2004).
- [2] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams *et al.*, *J. Mol. Graph. Model.* **19**, 26 (2001).
- [3] P. E. Wright and H. J. Dyson, *J. Mol. Biol.* **293**, 321 (1999).
- [4] A. L. Fink, *Curr. Opin. Struct. Biol.* **15**, 35 (2005).
- [5] V. N. Uversky and A. K. Dunker, *Biochem. Biophys. Acta* **1804**, 1231 (2010).
- [6] A. C. M. Ferreon, C. R. Moran, Y. Gambin, and A. A. Deniz, *Methods Enzymol.* **472**, 179 (2010).
- [7] V. N. Uversky, *Biochim. Biophys. Acta* **1834**, 932 (2013).
- [8] H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell Biol.* **6**, 197 (2005).
- [9] V. N. Uversky, *Protein Sci.* **11**, 739 (2002).
- [10] J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **84**, 7524 (1987).
- [11] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, *Science* **267**, 1619 (1995).
- [12] P. G. Wolynes, *Biochimie* **119**, 218 (2015).
- [13] K. Röder, J. A. Joseph, B. E. Husic, and D. J. Wales, *Adv. Theory Simul.* **2**, 1800175 (2019).
- [14] O. M. Becker and M. Karplus, *J. Chem. Phys.* **106**, 1495 (1997).
- [15] P. Garstecki, T. X. Hoang, and M. Cieplak, *Phys. Rev. E* **60**, 3219 (1999).
- [16] A. Caffish, *Curr. Opin. Struct. Biol.* **16**, 71 (2006).
- [17] D. J. Wales, *Curr. Opin. Struct. Biol.* **20**, 3 (2010).
- [18] M. Li, M. Duan, J. Fan, L. Han, and S. Huo, *J. Chem. Phys.* **139**, 185101 (2013).
- [19] M. Chwastyk, A. M. Vera, A. Galera-Prat, M. Gunnoo, D. Thompson, M. Carrion-Vazquez, and M. Cieplak, *J. Chem. Phys.* **147**, 105101 (2017).
- [20] K. Kloiber, R. Weiskirchen, B. Krautler, K. Bister, and R. Konrat, *J. Mol. Biol.* **292**, 893 (1999).
- [21] D. J. Wales and P. Salamon, *Proc. Natl. Acad. Sci. USA* **111**, 617 (2014).
- [22] A. B. Mantshyzov, A. S. Maltsev, J. Ying, Y. Shen, G. Hummer, and A. Bax, *Protein Sci.* **23**, 1275 (2014).
- [23] M. Chwastyk and M. Cieplak, *J. Phys. Chem. B* **124**, 11 (2020).
- [24] H. A. Lashuel, C. R. Overk, A. Oueslati, and E. Masliah, *Nat. Rev. Neurosci.* **14**, 38 (2013).
- [25] L. Stefanis, *Cold Spring Harb. Perspect. Med.* **2**, a009399 (2012).
- [26] Ł. Mioduszewski and M. Cieplak, *Phys. Chem. Chem. Phys.* **20**, 19057 (2018).
- [27] G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, and J. Mittal, *Proc. Natl. Acad. Sci. USA* **115**, 9929 (2018).
- [28] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, *J. Comput. Chem.* **26**, 1781 (2005).

- [29] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmueller, and A. D. MacKerell, Jr., *Nat. Methods* **14**, 71 (2017).
- [30] J. I. Sułkowska and M. Cieplak, *J. Phys.: Condens. Matter* **19**, 283201 (2007).
- [31] J. I. Sułkowska and M. Cieplak, *Biophys. J.* **95**, 3174 (2008).
- [32] K. Wołek, A. Gómez-Sicilia, and M. Cieplak, *J. Chem. Phys.* **143**, 243105 (2015).
- [33] The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrodinger, LLC, <http://www.pymol.org>.
- [34] E. R. Gansner and S. C. North, *Softw. Pract. Exper.* **30**, 1203 (2000).
- [35] N. J. Greenfield, *Nat. Protoc.* **1**, 2876 (2007).
- [36] M. J. O'Neil, *Anal. Chem.* **36**, 1238 (1964).
- [37] A. Sali, E. Shakhnovich, and M. Karplus, *Nature (London)* **369**, 248 (1994).
- [38] N. D. Succi and J. N. Onuchic, *J. Chem. Phys.* **101**, 1519 (1994).
- [39] M. Cieplak and J. R. Banavar, *Phys. Rev. E* **88**, 040702(R) (2013).
- [40] K. Wołek and M. Cieplak, *J. Chem. Phys.* **144**, 185102 (2016).
- [41] G. B. Akcapinar, A. Venturini, P. L. Martelli, R. Casadio, and U. O. Sezerman, *Prot. Eng. Des. Sel.* **28**, 127 (2015).
- [42] R. J. Allen, D. Frenkel, and P. R. ten Wolde, *J. Chem. Phys.* **124**, 024102 (2006).
- [43] A. K. Faradjan and R. Elber, *J. Chem. Phys.* **120**, 10880 (2004).
- [44] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, *J. Chem. Phys.* **108**, 1964 (1998).
- [45] B. W. Zhang, D. Jasnow, and D. M. Zuckerman, *J. Chem. Phys.* **132**, 054107 (2010).
- [46] G. A. Huber and S. Kim, *Biophys. J.* **70**, 97 (1996).
- [47] F. Noe, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weigl, *Proc. Natl. Acad. Sci. USA* **106**, 19011 (2009).
- [48] F. Sittel and G. Stock, *J. Chem. Phys.* **149**, 150901 (2018).
- [49] U. Sengupta, M. Carballo-Pacheco, and B. Strodel, *J. Chem. Phys.* **150**, 115101 (2019).
- [50] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, *Proteins* **17**, 412 (1993).
- [51] L. Molgedey and H. G. Schuster, *Phys. Rev. Lett.* **72**, 3634 (1994).
- [52] G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noe, *J. Chem. Phys.* **139**, 015102 (2013).
- [53] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Perez-Hernandez, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noe, *J. Chem. Theor. Comput.* **11**, 5525 (2015).
- [54] D. Shukla, C. X. Hernandez, J. K. Weber, and V. S. Pande, *Acc. Chem. Res.* **48**, 414 (2015).
- [55] B. E. Husic and V. S. Pande, *J. Am. Chem. Soc.* **140**, 2386 (2018).