Cite this: DOI: 10.1039/c4cp03402c

Circuit topology of self-interacting chains: implications for folding and unfolding dynamics†

Andrew Mugler,[abc] Sander J. Tans[c] and Alireza Mashaghi*[d]

Received 30th July 2014,
Accepted 8th September 2014

DOI: 10.1039/c4cp03402c

www.rsc.org/pccp

Understanding the relationship between molecular structure and folding is a central problem in disciplines ranging from biology to polymer physics and DNA origami. Topology can be a powerful tool to address this question. For a folded linear chain, the arrangement of intra-chain contacts is a topological property because rearranging the contacts requires discontinuous deformations. Conversely, the topology is preserved when continuously stretching the chain while maintaining the contact arrangement. Here we investigate how the folding and unfolding of linear chains with binary contacts is guided by the topology of contact arrangements. We formalize the topology by describing the relations between any two contacts in the structure, which for a linear chain can either be in parallel, in series, or crossing each other. We show that even when other determinants of folding rate such as contact order and size are kept constant, this 'circuit' topology determines folding kinetics. In particular, we find that the folding rate increases with the fractions of parallel and crossed relations. Moreover, we show how circuit topology constrains the conformational phase space explored during folding and unfolding: the number of forbidden unfolding transitions is found to increase with the fraction of parallel relations and to decrease with the fraction of series relations. Finally, we find that circuit topology influences whether distinct intermediate states are present, with crossed contacts being the key factor. The approach presented here can be more generally applied to questions on molecular dynamics, evolutionary biology, molecular engineering, and single-molecule biophysics.

1 Introduction

In mathematics, topology is the study of the properties of objects that are preserved through continuous deformations of the objects. A circle can be deformed into an ellipse by stretching, which implies that they are topologically equivalent. Two closed knots are topologically equivalent if they can be inter-converted by any deformation that does not include tearing. Focusing on an object's topology greatly reduces the amount of information one retains about its structure, since any property that can be continuously deformed, such as inter-object distances or geometric shape, is explicitly ignored.

In biology, structure plays a pivotal role in determining the functional properties of self-interacting molecular chains, such as proteins and ribonucleic acids. Indeed, a chain's structure, in terms of its intra-molecular contacts, is particularly important in determining the kinetics and pathways by which it folds and unfolds. For example, long-range contacts as well as stabilizing ones are known to guide folding of chains,[1–4] and the number of intra-molecular contacts is a key determinant of folding rates.[5] Contact order, defined as the average separation distance along the chain between contact sites, has also been shown to correlate well with folding rate.[6] Yet it remains unclear the extent to which folding properties are determined by structural features that are continuously deformable, such as inter-contact distances and overall chain geometry, or by the underlying topology of the chain, in the mathematical sense discussed above.

Defining the topology of self-interacting molecular chains in a mathematical sense requires careful distinction between continuous and discontinuous deformations. A large amount of previous work has appealed to the notion of topology to explore the properties of biomolecular chains. Some studies have used the term topology to describe physical features of a chain. This includes geometric properties, such as a chain's orientation with respect to surrounding structures[7–9] or properties of substructures within the chain,[9–11] *e.g.* the relative orientation of beta-strands and alpha-helices in proteins. This also includes distance-dependent properties, such as the set of all

[a] *Department of Physics, Purdue University, 525 Northwestern Avenue, West Lafayette, IN 47907, USA*
[b] *Department of Physics, Emory University, 400 Dowman Drive, Atlanta, GA 30322, USA*
[c] *FOM Institute AMOLF, Science Park 104, 1098 XG Amsterdam, The Netherlands*
[d] *Harvard Medical School, Harvard University, 25 Shattuck St, Boston, MA 02115, USA. E-mail: Alireza-Mashaghi-Tabari@meei.harvard.edu; Fax: +1-617-912-0117; Tel: +1-617-912-0256*

† Electronic supplementary information (ESI) available: Supplementary methods and results, including 4 figures. See DOI: 10.1039/c4cp03402c

inter-contact distances.[12–14] However, both geometry and distance are continuously deformable properties, and thus in this study we employ an alternative definition of topology.[15] Other studies have defined the topology of a chain in the context of knot theory,[16–21] which classifies chains based on the type of knots they form. However, knot theory generally ignores the arrangement of intra-molecular contacts. Contact arrangement is a topological property that we here need to include, since for a chain in its folded state, rearrangement requires breaking and reforming a contact, which is a discontinuous deformation.

Here we employ a mathematical definition of the topology of self-interacting chains that takes into account the connectivity established by intra-chain contacts.[15] This definition is invariant to inter-contact distances, is provably complete, and can be used in a principled way to determine structural equivalence. Because topology explicitly ignores contact details (*e.g.* their length and chemical nature) and is invariant to inter-contact distances, we can adopt a minimal model, which we then explore exhaustively. In our model, a chain contains specific sites that can each form a contact with one other site, leading to the formation of a fold with a specific topology. For any two contacts, one can recognize, by ignoring all other contacts, one of three possible arrangements, namely parallel, series, or cross (Fig. 1). Given the analogies with arrangements of electrical components in a circuit, we refer to this type of topology as "circuit topology". By classifying all the pairwise relations between contacts within a particular fold, the circuit topology of the fold is defined unambiguously.

Using this model, we study the effects of circuit topology on function, focusing specifically on the implications for the conformational dynamics of chains during folding and unfolding.

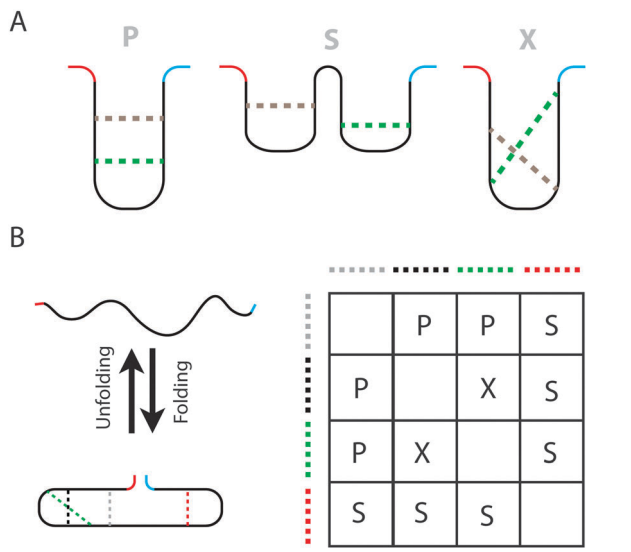

Fig. 1 Topology of a self-interacting molecular chain. (A) Any two contacts in the chain are either in a parallel (P), series (S), or cross (X) arrangement. To assign a relation to a contact pair, one can omit all the other contacts and determine the contact relation in the reduced system. Stretching of shrinking the chain does not change these relations. (B) A linear chain folds (unfolds) by forming (breaking) intramolecular contacts. The topology of the folded chain can be represented by a matrix whose elements describe all inter-contact relations.

Folding here is the sequential formation of contacts between two sites on the chain, while unfolding is the sequential disruption of contacts upon pulling the chain ends apart. We find that the circuit topology (hereafter referred to simply as topology) imposes fundamental constraints on the conformational search during both folding and unfolding. Topology sets selection rules for unfolding transitions, forbidding some transitions while permitting others, with the number of forbidden transitions increasing with the fraction of parallel relations and decreasing with the fraction of series relations. We further show that topology is a generic determinant of folding rate. In particular, we demonstrate that the impact of topology goes beyond distance-based measures like contact order, in the sense that two chains with the same contact order but different topologies can have markedly different folding rates. Indeed, folding rate increases with the fractions of parallel and crossed contacts. Finally, we demonstrate that topology is a key determinant of the number of intermediate folding and unfolding states. The presence of crossed contacts emerges as a necessary ingredient for observing several distinct intermediate states.

2 Results

2.1 Defining topology

Our definition of topology[15] is based on the realization that any two pairs of contacts between sites on a chain must exist in one of three arrangements: contacts are either (i) parallel to each other, (ii) in series with each other, or (iii) crossed (Fig. 1A). Indeed, even when describing chains in terms of inter-contact distances, these three categories emerge naturally and unavoidably.[22,23] This categorization is complete, in the sense that even for molecular chains with many contacts, every pair of contacts must be related in one of these three ways.[15] The relations between contact pairs therefore completely specify the topology of any linear chain. The topology is shown in matrix form in Fig. 1B. Any two chains with the same topology matrix are topologically equivalent. Importantly, this definition is invariant to the actual distances between adjacent contact sites; the topology remains unchanged if these distances stretch or shrink.

To investigate the role of topology in determining functional properties, we consider all linear molecular chains with N monovalent contact sites and C binary contacts. We first focus on how topology affects the dynamics of unfolding upon stretching the two ends of the chain, then the dynamics of folding, and finally the properties of the configurational space explored during both folding and unfolding.

2.2 Topology correlates with forbidden transitions during unfolding

To understand the constraints that topology places on unfolding, we consider a chain that is pulled apart at the two ends. We define a single step in the unfolding process as the breaking of one contact. We assume that only the contacts lying along the shortest path(s) between the chain ends experience tension from the pulling, and therefore we specify that only these

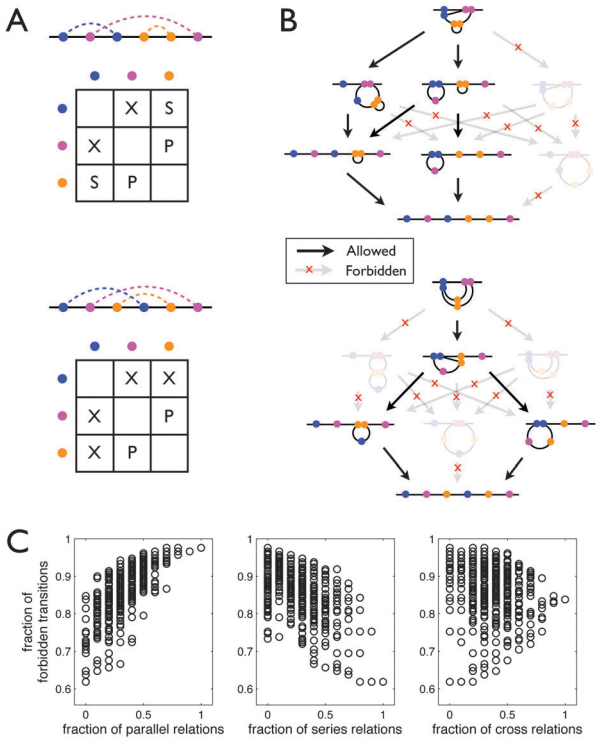

Fig. 2 Topology shapes unfolding dynamics by forbidding transitions. (A) Two chains with $C = N/2 = 3$ contacts are shown (in their unfolded states), along with their topology matrices. (B) Unfolding trees corresponding to each chain are shown, revealing all possible unfolding trajectories. Many transitions are forbidden, resulting in two different tree structures. (C) The fraction of forbidden transitions is plotted against the fractions of each topological relation for all 513 topologically unique chains with $C = N/2 = 5$ contacts.

contacts may break in each step. Finally, we assume that broken contacts may not reform. This produces for each chain a "tree" of possible unfolding trajectories (Fig. 2). Along each trajectory, a total of $C + 1$ states, including the fully folded and fully unfolded states, will be observed as contacts break one by one. A similar force-dependent, sequential unfolding model has been used previously to probe the structure-dependent properties of two proteins;[24] here we use such a model to investigate the effects of topology across a wide class of linear chains.

Fig. 2A shows two example chains with $C = N/2 = 3$, and Fig. 2B shows their corresponding unfolding trees (an unfolding trajectory is also depicted in terms of topology matrices in the ESI† Appendix). One observes in Fig. 2B that, for a given chain, while some trajectories are allowed, others are forbidden. This is a topological effect: when the chain is pulled at its ends, only a subset of its contacts, as determined by the topology of the chain, lie along the shortest path and experience the tension necessary to break. This subset then determines which unfolding transitions are allowed and which are forbidden under the pulling protocol. Moreover, Fig. 2B also shows that the impact of topology on which transitions are forbidden can be pronounced. The two chains in Fig. 2A differ only in the fact that two neighboring contact sites have been swapped. Yet, this swap changes the topology (as seen in the matrices in Fig. 2A), and leads to a large change in the structure of the trees (Fig. 2B). This suggests that small changes in contact arrangement can lead to large changes in the unfolding dynamics.

To quantify the extent to which topology influences unfolding, we summarize a given chain's topology by the relative fractions of parallel, series, and cross relations. We then compare this fraction to the fraction of forbidden transitions in its unfolding tree. Fig. 2C shows this comparison for every chain with $C = N/2 = 5$. Two trends are immediately apparent. First, the fraction of forbidden transitions increases with the fraction of parallel relations (left plot). This is logical because in a parallel relation, the outer contact shields the inner contact from the applied tension (Fig. 1A). Indeed, the unfolding trajectory of a chain with only parallel relations has only one path, namely the path in which contacts are "unzipped" one by one. As a result, increasing the fraction of parallel relations in a folded chain's topology generally increases the fraction of forbidden transitions in its unfolding tree, as seen in Fig. 1C.

The second trend apparent in Fig. 2C is that the fraction of forbidden transitions decreases with the fraction of series relations (middle plot). Unlike in a parallel relation, in a series relation each contact is independent of the other (Fig. 1A). Therefore, neither pathway—one contact breaking first or the other breaking first—is forbidden. For this reason, increasing the fraction of series relations in a chain's topology generally decreases the fraction of forbidden transitions in its folding tree, as seen in Fig. 2C.

In a cross relation, neither contact shields the other from tension, but the two contacts are not independent either (Fig. 1A). Thus we might expect that the two competing effects seen for parallel and series relations would balance to a certain degree. Indeed, Fig. 2C shows that the fraction of forbidden transitions neither strongly increases nor strongly decreases with the fraction of cross relations (right plot). Instead, the most noticeable effect is a tapering of the data as the fraction of cross relations increases. The tapering is a consequence of the "density of states" of chains in the space of contact relations: there are very few chains with a fraction of cross relations near 1 (indeed, there is only one chain with this fraction equal to 1); whereas there are many chains with zero cross relations, since there are many ways to have a mixture of parallel and series relations. This effect is not unique to cross, and indeed a similar tapering is present in the left and middle plots of Fig. 2C, convolved with the general trends. Although Fig. 2C illustrates that the fraction of cross relations does not have a strong effect on forbidden transitions, we will see below that cross relations are an important determinant of other properties during folding and unfolding.

Finally, we note that forbidden transitions may arise in an unfolding tree in several distinct ways: (i) the contact might not lie along the shortest path(s), (ii) the transition might originate from an inaccessible state, or (iii) the transition might involve breaking two contacts and reforming another. Topology affects all of these types of forbidden transitions. The mechanisms

are similar to those underlying Fig. 2C and are discussed in the ESI† Appendix.

2.3 Topology guides folding dynamics and affects folding rate

Past research has shown that the folding rate of small proteins strongly correlates with their contact order, which is defined as the mean inter-contact distance divided by the protein length.[6] For large proteins an important determinant is the size.[25] What determines the folding rate is still incompletely understood. However there is an agreement that the folding rate correlates with the metrics of native fold.[14,26] Here we investigate how topology guides folding dynamics and how folding rate and topology are related.

To understand the constraints that topology places on folding, we model folding as a step-wise memoryless process.[27] The time for each contact to form is proportional to the shortest inter-contact distance along the partially folded chain, raised to the 3/2 power, a scaling that is predicted analytically[28,29] and has been measured for proteins.[30,31] We consider both deterministic folding, in which the contact with the shortest formation time forms first, and stochastic folding, in which the probability of a contact forming is inversely proportional to its formation time. The folding rate is then calculated as the inverse of the average total time to go from the fully unfolded state to the fully folded state. Results for deterministic folding are shown in Fig. 3. Stochastic folding gives similar results (ESI† Appendix).

Fig. 3A shows folding trees for the simplest case of chains with just two contacts. Above each tree is the folding rate, assuming that the most probable folding path is taken (deterministic folding). The chain folds fastest in the parallel case, followed by the cross, and then the series. The reason is topological: in the parallel arrangement, formation of the inner contact reduces the formation time of the outer contact by bringing its contact sites closer together, effectively allowing the chain to "zip up" into its folded state. The cross arrangement features partial nesting of contacts, and therefore also allows for a speedup due to zipping, albeit to a lesser extent. The series arrangement, however, does not involve any nesting of contacts—indeed the formation of each contact does not affect the folding time of the other—and therefore the series relation does not result in any speedup.

Experimentally, the folding rates of small proteins are known to decrease with their contact order.[6] Fig. 3B shows that in our model the folding rate indeed decreases with contact order. However, all three chains in Fig. 3A have the same contact order, and yet they have different folding rates. This suggests that topological features other than contact order can affect the folding rate. To systematically explore the effects of topology on folding rate, we consider a group of chains with the same contact order by fixing the inter-contact distances while varying the contact positions. To combine results from many such groups, the folding rate r for each chain is normalized by the minimum rate r_0 for the group, which occurs when all contacts are in series. The results of this procedure are shown in Fig. 3C and D.

Fig. 3C reveals that the folding rate increases as the fraction of parallel relations increases. Indeed, a chain with all parallel

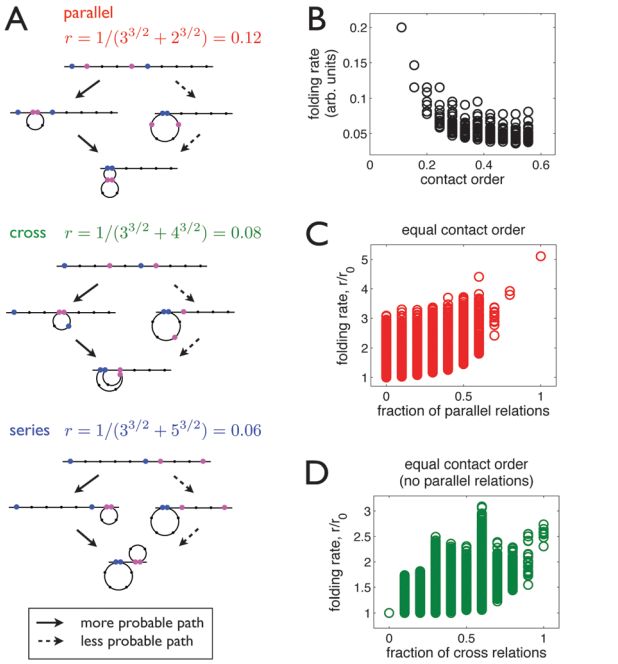

Fig. 3 Topology affects the folding rate. (A) Folding trees for three chains, each with two contacts in parallel (top), cross (middle), or series (bottom). All have the same contact order, but the parallel case folds fastest, followed by the cross, then the series. Folding rate (in arbitrary units) assumes the more probable path is taken (deterministic folding). (B) Folding rate decreases with contact order. Data are for all 513 topologically unique chains with length $N = 10$ and $C = 5$ contacts. (C) At constant contact order, folding rate increases with the fraction of parallel relations. (D) At constant contact order, for chains with no parallel relations, folding rate increases with the fraction of cross relations. In (C) and (D), data are for 10^6 chains with $C = 5$ contacts in randomly sampled arrangements and N sufficiently large to accommodate the all-series arrangement, whose folding rate is r_0.

relations achieves more than a 5-fold speedup compared to the all-series arrangement. This speedup is due to the zipping effect discussed above. The fact that the folding rate increases continually with the fraction of parallel relations indicates that the effect is more general than the image of a perfect "zipper" implies: not all contacts have to be in parallel in order to achieve a speedup in the folding rate. Moreover, since parallel arrangement is a topological property and not a physical one, the trend in Fig. 3C shows that the parallel contacts are not required to be geometrical neighbors in order to see the advantage of the zipping effect on the folding rate. Note that this topological advantage occurs even when the contact order is kept constant.

Fig. 3D reveals that the folding rate also increases with the fraction of cross relations. Indeed, a chain with all cross relations achieves nearly triple the folding rate compared to the all-series arrangement. We note that Fig. 3D includes only chains with no parallel relations. The reason is that we expect an increase in cross relations to impart a speedup so long as it does not come at the expense of a decrease in parallel relations, since the latter is expected to have the stronger effect on folding rate (Fig. 3A). The increase seen in Fig. 3D therefore illustrates that, beyond the speedup seen for parallel contacts, which are fully nested, the advantage of zipping also extends to crossed contacts, which are partially nested.

Finally, we point out that cross relations impart a further geometric folding advantage that is not captured in Fig. 3D, but is suggested pictorially by the more probable paths in Fig. 3A. In our model, the folding time for a contact is based on the shortest inter-contact distance along the chain. In the case of parallel and series relations, this distance, which is inherently one-dimensional, may be an accurate proxy for the actual inter-contact distance in the chain's three-dimensional geometry. However, in the case of a cross relation, the fact that the contacts are crossed means that formation of the first contact brings the second contact's sites closer together in three-dimensional space than this one-dimensional distance would imply. We expect that this geometrical effect would cause a further increase in the folding rate with the fraction of cross relations, beyond that seen in Fig. 3D.

2.4 Topology drives two- or multi-state folding and unfolding

Kinetic models with two or more states[32,33] have been the dominant paradigm for understanding protein folding dynamics. Multi-state folding and unfolding is also important for bio-molecular engineering,[34] and is exploited in designing molecular switches. What determines if a molecule exhibits two-state or multi-state folding or unfolding? The answer to this question is still unclear. Here we show that topology is a determinant for the presence of (un)folding intermediates.

As a chain folds or unfolds, it takes on a number of intermediate configurations, as shown by the trees in Fig. 2B and 3A. Each of these configurations has a characteristic size, which we define to be the length of the shortest path from one end of the chain to the other. In principle, the length along any of our idealized chains, measured as the number of inter-site segments in the shortest path, may differ from the physical length of a molecular chain with that topology. For this reason we restrict the analysis here to chains in which every site participates in a contact (Fig. 4), which makes length equivalent to a topological property. Averaging over all possible transition paths, while weighting by the probability of observing each path (for details see the ESI† Appendix), we obtain a mean number of times each length is visited during either the folding or the unfolding process. Since some configurations are the same length (*i.e.* that length is degenerate), this number can be greater than one. We therefore call this number the path-weighted degeneracy.

Fig. 4A–D shows the path-weighted degeneracy for chains with all series relations, all parallel relations, all cross relations, and a mix of parallel and cross. The all-series and all-parallel cases are straightforward to understand. In the all-series case (Fig. 4A), we see that all intermediate lengths are visited once, no matter what path is taken. This is because contacts in series form or break independently from each other, allowing the chain to fold or unfold in a continuous fashion, visiting all possible lengths in between. In the all-parallel case (Fig. 4B), most even lengths are visited, while most odd lengths are not. This is because contacts in parallel are nested, such that the inner contact forms (or the outer contact breaks) with the highest probability in each step, and both of these events lead to a change in length of two units. Folding and unfolding with

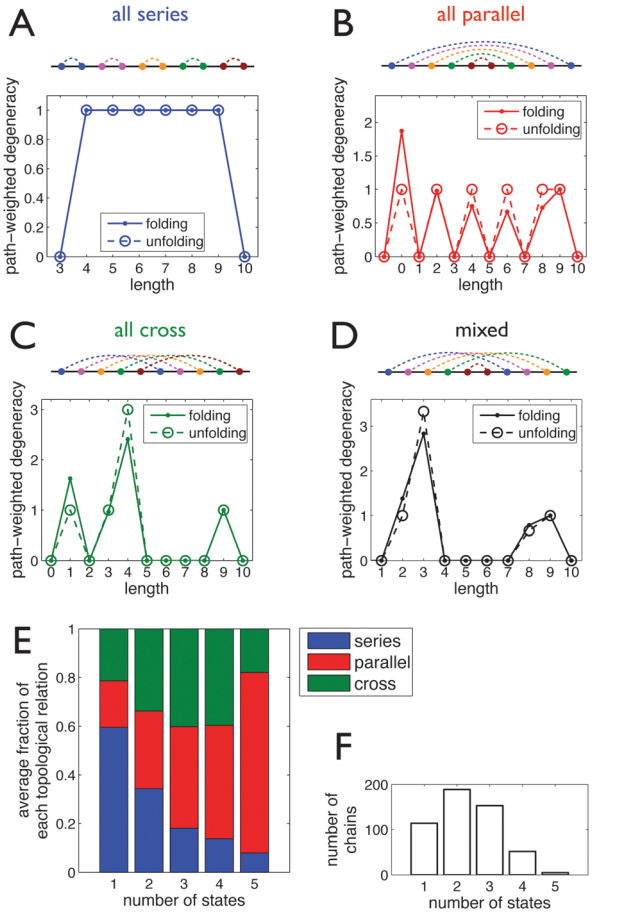

Fig. 4 Topology influences the number of intermediate states visited during folding or unfolding. Example chains are shown with (A) all series relations, (B) all parallel relations, (C) all cross relations, and (D) a mix of 60% cross and 40% parallel. In each case, for both folding and unfolding, we plot the number of times each molecular length is visited, averaged over all possible paths (the "path-weighted degeneracy"). States are groups of one or more consecutive lengths separated by unvisited lengths, such that A–D have 1, 5, 3, and 2 states, respectively. For all chains with length $N = 10$ and $C = 5$ contacts that exhibit a given number of states, we plot (E) the average fractions of each topological relation and (F) the number of such chains out of the total of 513.

all parallel contacts is therefore a quasi-continuous process, similar to the case of all series.

The picture is different in the all-cross case (Fig. 4C). Here, three distinct lengths or groups of consecutive lengths are visited, forming three distinct peaks in the path-weighted degeneracy plot. Two of the peaks are well-separated: between the second and third peak in Fig. 4C, there are four lengths that are never visited. Distinct peaks arise because cross relations are highly interconnected, leading to the requirement that several contacts form or break before the length changes appreciably. The large peak separation arises because when these several contacts ultimately do form or break, the subsequent length change can be large. Therefore, unlike series or parallel relations, which lead to a largely continuous exploration of intermediate lengths, Fig. 4C suggests that cross relations give rise to length distributions that are concentrated and discrete. Distinct, well-separated peaks are

also possible when cross relations are mixed with other topological relations: Fig. 4D shows an example of a chain with 60% cross relations and 40% parallel relations, that exhibits two clearly separated peaks.

What is the typical topological composition of chains with a given number of intermediate states? We answer this question by defining intermediate states as peaks in the path-weighted degeneracy plot. That is, we define states as consecutive groups of lengths that are visited with nonzero probability (the chains in Fig. 4A–D have 1, 5, 3, and 2 states, respectively, under this definition). To address the question systematically, we consider all chains of length $N = 10$ with $C = 5$ contacts. For all such chains that exhibit a given number of states, we find the average fractions of series, parallel, and cross relations. Fig. 4E and F show the results of this procedure for the case of unfolding. Results for folding are almost identical (ESI† Appendix). Looking at Fig. 4E, three key features emerge.

First, as seen in the leftmost bar in Fig. 4E, chains that exhibit only one state are dominated by series relations. Indeed, as the number of states increases from one, the fraction of series relations decreases. This is a general consequence of the effect seen in Fig. 4A: series relations allow for continuous folding and unfolding, allowing a chain to access many consecutive lengths, which together form one continuous state.

Second, as seen in the rightmost bar in Fig. 4E, chains that exhibit the maximum of five states are dominated by parallel relations. Indeed, as the number of states decreases from five, the fraction of parallel relations decreases. This is a general consequence of the effect seen in Fig. 4B: parallel relations allow for quasi-continuous folding and unfolding, allowing a chain to access every two lengths, leading to many finely separated states.

The third key feature of Fig. 4E is that to observe an intermediate number of states (2, 3, or 4 states), a sizable fraction of cross relations is essential. Indeed, as a function of the number of states, the fraction of cross relations reaches a maximum at three states. This is a general consequence of the effect seen in Fig. 4C: cross relations yield highly interconnected chains that undergo appreciable length changes in several discrete jumps. Importantly, on average the fraction of cross relations is not dominant in these chains; rather, intermediate state numbers emerge when cross relations coexist with parallel relations in roughly equal proportion, similar to the example in Fig. 4D. This implies that cross relations are necessary to produce several discrete folding or unfolding states, but that not all topological relations need to be cross. Finally, we note that these intermediate state numbers are particularly important because, as seen in Fig. 4F, chains with two or three states are the most common of the group.

3 Discussion and conclusion

Here we employed a topological description of self-interacting linear chains beyond what is offered by knot theory and the metrics of native fold such as contact order. The introduced topology of contact arrangements, termed circuit topology, is not only a descriptor of the molecular shape but also is a determinant of its folding and unfolding dynamics. Our study shows that topology guides the conformational search and sets fundamental constraints on conformational transitions.

We have shown that topology can constrain the folding and unfolding dynamics of a molecular chain. In particular, we have found that the number of forbidden unfolding transitions increases with the fraction of parallel relations, thus rendering the unfolding dynamics more deterministic. Moreover, we have found that the folding rate increases with the fractions of parallel and crossed relations. Topology therefore complements standard determinants of folding kinetics and hence should be considered alongside energetic and geometric measures in predicting the dynamics of molecular chains. In some cases, considering topology alone and considering a different measure alone, such as bond energy, can lead to contrasting predictions (a simple example is illustrated in the ESI† Appendix). In such cases, it is particularly important to consider multiple measures in conjunction to develop the most accurate predictive framework.

We have focused only on ideal chains with binary contacts. However, the theory and the computational protocol is general and can be extended to chains with multivalent contacts. In the ESI† Appendix we summarize in the form of flowcharts the computational protocol used to produce the numerical data in Fig. 2–4, to elucidate the logic of the procedure and facilitate future extensions.

Our study also yields testable predictions. Comparison to nucleic acid and protein conformational dynamics can be made as long as the crystal structure or contacting residues are known. Contacts can be defined in different ways, depending on the molecule and question of interest. For RNA, base pairing is a natural choice, while in proteins beta–beta contacts or residue contacts are appropriate options. Contacts can also be defined not only based on structure but also by considering the associated energies. For instance, contacts that are not sufficiently tight could be ignored. One can then ask if the folding rate of proteins or RNAs correlates with, for example, the fraction of parallel relations. Another way to test our predictions is to design and synthesize molecules with desired topological properties. For example, two RNA molecules (or two synthetic polymers) with identical lengths and contact orders, can be designed with two or more contacts being in series or parallel. Interrogation of these synthetic systems with optical tweezers/AFM would allow one to experimentally test the predictions developed herein. The single-molecule force methods allow for direct measurement of folding rate of single bimolecular chains as well as their folding and unfolding pathways.[35,36]

Our results on the implications of topology for unfolding are based on force-induced unfolding of molecules *via* pulling at the two ends. Examples of pulling-induced unfolding are seen in nature.[37,38] Alternatively, molecules can be unfolded thermally or chemically, processes that are mechanistically very different from pulling.[39] The unfolding pathways available for the molecule are then set by its topology, geometry, and energetics, as well as the applied unfolding mechanism. Our approach may also find practical applications in single-molecule pulling studies, where it can be used to infer structural information from observed unfolded lengths.

We provided evidence that the topology of the chain sets the size and number of intermediate states during folding and unfolding. Our results on the (un)folding intermediates may be of biological significance, as the physiological role of many proteins relies on their fold size. The fold size determines the propensity of proteins for cleavage,[40,41] the translocation efficiency,[42] and the binding affinity in physiological[43] and pathological[44] conditions. The fact that the arrangement of contacts could give rise to a certain size distribution is also important from an evolutionary biology perspective,[45] as well as for molecular engineering.[46] Biomolecules are evolved to have a high designability and a large tolerance to changes in primary sequence.[47] Nonetheless, the design of multi-state protein-based molecular switches turns out to be a challenge for protein engineering.[47] Our finding that topology influences the multiplicity of intermediate states of self-interacting chains, along with advances in polymer chemistry that have allowed synthesis of chains with a desired arrangement of contacts,[48–52] holds promise for future rational design of molecules with new functions.

Finally, it would be interesting to study whether proteins (or nucleic acids) with similar topology also share similar biological function, as well as the extent to which topology has been evolutionarily conserved.

Acknowledgements

The authors thank Sanne Abeln for critical reading of the manuscript. SJT was supported by the research programme of the Foundation for Fundamental Research on Matter (FOM), which is part of the Netherlands Organisation for Scientific Research (NWO).

References

1 M. Vendruscolo, E. Paci, C. M. Dobson and M. Karplus, *Nature*, 2001, **409**, 641–645.
2 S. W. Lockless and R. Ranganathan, *Science*, 1999, **286**, 295–299.
3 N. Go and H. Taketomi, *Proc. Natl. Acad. Sci. U. S. A.*, 1978, **75**, 559–563.
4 W. Meng, N. Lyle, B. Luan, D. P. Raleigh and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 2123–2128.
5 D. E. Makarov, C. A. Keller, K. W. Plaxco and H. Metiu, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 3535–3539.
6 D. Baker, *Nature*, 2000, **405**, 39–42.
7 C. Manoil and J. Beckwith, *Science*, 1986, **233**, 1403–1408.
8 M. Rapp, E. Granseth, S. Seppälä and G. Von Heijne, *Nat. Struct. Mol. Biol.*, 2006, **13**, 112–116.
9 G. von Heijne, *Nat. Rev. Mol. Cell Biol.*, 2006, **7**, 909–918.
10 G. MacBeath, P. Kast and D. Hilvert, *Science*, 1998, **279**, 1958–1961.
11 D. Shortle and M. S. Ackerman, *Science*, 2001, **293**, 487–489.
12 L. Holm and C. Sander, *et al.*, *J. Mol. Biol.*, 1993, **233**, 123–138.
13 Y. Zhang and J. Skolnick, *Nucleic Acids Res.*, 2005, **33**, 2302–2309.
14 M. Rustad and K. Ghosh, *J. Chem. Phys.*, 2012, **137**, 205104.
15 A. Mashaghi, R. J. van Wijk and S. J. Tans, *Structure*, 2014, **22**, 1227–1237.
16 S. A. Wasserman and N. R. Cozzarelli, *Science*, 1986, **232**, 951–960.
17 A. L. Mallam and S. E. Jackson, *Nat. Chem. Biol.*, 2011, **8**, 147–153.
18 S. Wallin, K. B. Zeldovich and E. I. Shakhnovich, *J. Mol. Biol.*, 2007, **368**, 884–893.
19 A. R. Mohazab and S. S. Plotkin, *PLoS One*, 2013, **8**, e53642.
20 J. Arsuaga, M. Vazquez, P. McGuirk, S. Trigueros, D. W. Sumners and J. Roca, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 9165–9169.
21 J. I. Sukowska, E. J. Rawdon, K. C. Millett, J. N. Onuchic and A. Stasiak, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, E1715–E1723.
22 M. Muthukumar and B. G. Nickel, *J. Chem. Phys.*, 1984, **80**, 5839–5850.
23 K. M. Fiebig and K. A. Dill, *J. Chem. Phys.*, 1993, **98**, 3475–3487.
24 D. K. Klimov and D. Thirumalai, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 7254–7259.
25 D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker and A. V. Finkelstein, *Protein Sci.*, 2003, **12**, 2057–2062.
26 P. F. Fasca, R. D. Travasso, A. Parisi and A. Rey, *PLoS One*, 2012, **7**, e35599.
27 Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.*, 1987, **84**, 6611–6615.
28 A. Szabo, K. Schulten and Z. Schulten, *J. Chem. Phys.*, 1980, 4350.
29 T. Guerin, O. Benichou and R. Voituriez, *Nat. Chem.*, 2012, **4**, 568–573.
30 O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello and T. Kiefhaber, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 9597–9601.
31 L. J. Lapidus, W. A. Eaton and J. Hofrichter, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 7220–7225.
32 P. Kim and R. Baldwin, *Annu. Rev. Biochem.*, 1990, **59**, 631–660.
33 K. A. Beauchamp, R. McGibbon, Y.-S. Lin and V. S. Pande, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 17807–17813.
34 K. Inaba, N. Kobayashi and A. Fersht, *J. Mol. Biol.*, 2000, **302**, 219–233.
35 A. Mashaghi, S. Mashaghi and S. J. Tans, *Angew. Chem., Int. Ed.*, 2014, DOI: 10.1002/anie.201405566.
36 P. Bechtluft, R. G. H. van Leeuwen, M. Tyreman, D. Tomkiewicz, N. Nouwen, H. L. Tepper, A. J. M. Driessen and S. J. Tans, *Science*, 2007, **318**, 1458–1461.
37 E. A. Evans and D. A. Calderwood, *Science*, 2007, **316**, 1148–1153.
38 C. P. Johnson, H.-Y. Tang, C. Carag, D. W. Speicher and D. E. Discher, *Science*, 2007, **317**, 663–666.
39 G. Stirnemann, S.-g. Kang, R. Zhou and B. J. Berne, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 3413–3418.
40 X. Zhang, K. Halvorsen, C.-Z. Zhang, W. P. Wong and T. A. Springer, *Science*, 2009, **324**, 1330–1334.

41 A. J. Jakobi, A. Mashaghi, S. J. Tans and E. G. Huizinga, *Nat. Commun.*, 2011, **2**, 385.
42 K. R. Mahendran, M. Romero-Ruiz, A. Schlösinger, M. Winterhalter and S. Nussberger, *Biophys. J.*, 2012, **102**, 39–47.
43 V. Vogel, *Annu. Rev. Biophys. Biomol. Struct.*, 2006, **35**, 459–488.
44 A. Ahmad, I. S. Millett, S. Doniach, V. N. Uversky and A. L. Fink, *Biochemistry*, 2003, **42**, 11404–11416.
45 C. Debès, M. Wang, G. Caetano-Anollés and F. Gräter, *PLoS Comput. Biol.*, 2013, **9**, e1002861.
46 S. Seetharaman, M. Zivarts, N. Sudarsan and R. R. Breaker, *Nat. Biotechnol.*, 2001, **19**, 336–341.
47 D. Shortle, K. T. Simons and B. David, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 11158–11162.
48 K. L. Wooley, J. S. Moore, C. Wu and Y. Yang, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 11147–11148.
49 P. Englebienne, P. A. Hilbers, E. Meijer, T. F. De Greef and A. J. Markvoort, *Soft Matter*, 2012, **8**, 7610–7616.
50 B. T. Tuten, D. Chao, C. K. Lyon and E. B. Berda, *Polym. Chem.*, 2012, **3**, 3068–3071.
51 E. Appel, J. Dyson, J. del Barrio, Z. Walsh and O. Scherman, *Angew. Chem., Int. Ed.*, 2012, **51**, 4185–4189.
52 T. Mes, R. van der Weegen, A. R. Palmans and E. Meijer, *Angew. Chem., Int. Ed.*, 2011, **50**, 5085–5089.

