Circuit Topology of Proteins and Nucleic Acids

Alireza Mashaghi*, Roeland J. van Wijk, and Sander J. Tans
FOM institute AMOLF, Science Park 104, 1098 XG Amsterdam, the Netherlands.

Folded biomolecules display a bewildering structural complexity and diversity(Onuchic and Wolynes, 2004; Sali et al., 1994). They have therefore been analyzed in terms of generic topological features(Bailor et al., 2010; Baker, 2000; Chang and Tinoco, 1994; Li et al., 2006; Melchers et al., 1997; Meyer, 2000; Richardson, 1981). For instance, folded proteins may be knotted(Taylor, 2000), have beta-strands arranged into a greek-key motif(Hutchinson and Thornton, 1993), or display high contact order(Baker, 2000). In this perspective, we present a method to formally describe the topology of all folded linear chains, and hence provide a general classification and analysis framework for a range of biomolecules. Moreover, by identifying the fundamental rules that intra-chain contacts must obey, the method establishes the topological constraints of folded linear chains. We also briefly illustrate how this circuit topology notion can be applied to study the equivalence of folded chains, the engineering of artificial RNA structures and DNA origami, the topological structure of genomes, and the role of topology in protein folding.

1 Current topology approaches

Topology is a mathematical notion that has been used to describe properties of objects that remain unchanged under a certain kind of continuous, invertible and one-to-one transformation(Mizuguchi and Go, 1995). Examples of such transformations include bending, stretching and shrinking. Objects like squares, circles and triangles are inter-convertible by such transformations and therefore belong to the same topological class. Instead, the “figure 8” knot is not inter-convertible to any of these three objects unless a connection is torn apart (Fig. 1a).

In chemistry, topology is a convenient way of describing elementary features of the structure of molecules(Brown, 2002; Flapan, 2000; Li et al., 2006). This is important not only for classification, but also because the structure and shape of a molecule sets many of its properties(Flapan, 2000; Liang and Mislow, 1995; Yamamoto, 2012). Moreover, topology is relevant to engineering(Ayme et al., 2012; Blankenship and Dawson, 2007; Coskun et al., 2012; Harada, 2012; Yan et al., 2002). Engineered molecules with complex topologies may display emergent properties(Kamien, 2003; Siegel, 2004), and topological classifications can provide guidelines for chemical synthesis(Guan et al., 1999; Tezuka and Oike, 2001).

In biology, molecular structures are astonishingly complex and diverse. At the same time, certain structural features can be highly conserved. For instance, the mammalian metabolic enzyme glycogen phosphorylase, was found to contain a structural core similar to the T4 phage DNA glucosyltransferase(Pauling and Corey, 1951a; Sibanda and Thornton, 1991; Wetlaufe.Db, 1973), a protein that almost cannot be more distant in terms of function and taxonomy. Understanding the diversity of biomolecular structures and its functional consequences is considered one of the key scientific challenges in biology(Holm and Sander, 1996). Proteins have for instance been classified by visual inspection(Richardson, 1981), geometry(Holm and Sander, 1996), the nature of transition intermediates(Milner-White and Poet, 1986), and the spatial arrangement of secondary structures(Moutevelis and Woolfson, 2009; Taylor, 2002). The notion of topology could present a powerful tool to address this issue, as it has been shown to yield unifying structural relationships among apparently diverse molecules and more complex materials(Sabato, 1970; Senyuk et al., 2013; Terentjev, 2013). Moreover, topologies of RNA and chromosome structures can have important functional implications and are dramatically altered in many diseases(Bailor et al., 2010; Cavalli and Misteli, 2013).

Let us consider three important topological notions that have been introduced so far: branch topology, knot topology, and network topology. Note that the term topology is sometimes confused with geometry(Francl, 2009). Occasionally it is used to refer to the molecule's orientation with respect to surrounding structures(Manoil and Beckwith, 1986; Rapp et al., 2006; von Heijne, 2006), to intramolecular chain orientation(MacBeath et al., 1998; Shortle and Ackerman, 2001), to the number and proximity of secondary structural elements within the protein(Li et al., 2006; MacBeath et al., 1998; Meyer, 2000), or to describe permutations in primary sequence (Shank et al., 2010). Here we use the term topology in the mathematical sense (Fig. 1a).

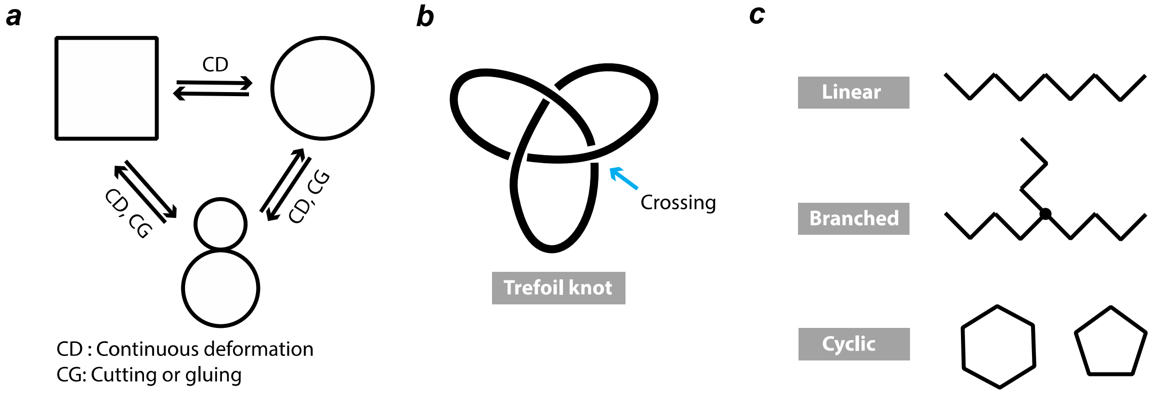


Figure 1: Topological polymer chemistry. (a) Continuous deformation (CD) changes a shape to q topologically equivalent shape. Cutting and gluing (CG) operation can change a topology. The topology of circles and rectangles are identical but different from the figure 8. (b) The trefoil knot is a knot topology seen in RNA methyltransferase. The trefoil knot has three crossings in its minimal representation. (c) Hydrocarbons can be classified by their topologies, such as linear, branched and cyclic. Multiple topological invariants can be identified such as the total number of chain ends and of branch points(Tezuka and Oike, 2001).

Within branch topology, one can distinguish linear, branched and cyclic topologies (Tezuka and Oike, 2001) (Fig. 1c). Here, a number of properties are invariant under continuous deformations: the total number of chain ends (termini); the total number of branch points (junctions); the number of branches at each junction and the connectivity of the junction. All linear polymers, and hence also all folded proteins and RNA, belong to one topological class and thus are topologically identical.

In knot topology, a central issue is to identify the topological features related to knots. We know intuitively that upon stretching a rope with a knot, certain features remain identical. These features are referred to as the topology of the knot. In contrast, the end-to-end distance can change upon stretching and is thus not a topological property. Pulling operations have therefore been used to identifying topological features, which is also referred to as the recognition problem. Topological knots are created by starting with a linear chain, wrapping it around itself to form a physical knot, and then fusing its two free ends together to form a closed loop (Fig. 1b). Such closed knots are equivalent if and only if they can be interconverted by stretching and twisting(Kauffman, 1994). A closed knot can be represented with its projection on a plane, the so called knot diagram. Characterizing the topology of a knot is not straightforward. Briefly, for a given knot there is a projection that minimizes the number of chain crossing. As these crossings cannot be changed without tearing or gluing the chain, they can characterize the topology. For example, two topologically equivalent knots must have equal number of crossings. The reverse statement does not necessarily hold true. Further characterization of the topology (e.g. by analyzing the mirror image and braids) is needed for unique identification.

Molecular knots occur naturally in biological systems and have been engineered for technological applications. Controlled synthesis of molecular knots has recently been made possible: organic molecules have been shown to self-assemble into closed trefoil knots, in a process driven by hydrophobic interactions (Ponnuswamy et al., 2012; Siegel et al., 2012). At room temperature, sufficiently long (Sumners and Whittington, 1988) linear equilibrated polyelectrolyte chains such as DNA molecules show self-entanglement and can form physical knots, which may for instance complicate replication(Lopez et al., 2012) or prohibit translocation through pores (Rosa, 2012). Additional functional consequences of biomolecular knot formation are kinetic(Soler and Faisca, 2013) and structural stability(Mallam et al., 2010; Sulkowska et al., 2008) of the knotted conformations. Physical knots are extremely rare in RNA molecules (VanLoock et al., 1998). It has been shown that polypeptide chains can self-entangle and form knots (Mallam et al., 2010; Noel et al., 2013; Skrbic et al., 2012). However, only a small fraction of proteins (<1%) in the Protein Data Bank, including rRNA methyltransferases, carbonic anhydrases and ubiquitin hydrolase, are identified as knotted (Sulkowska et al., 2012; Virnau et al., 2006). Overall, knot topology is therefore also not very useful when distinguishing RNA or protein folds, as most would again fall in the same class.

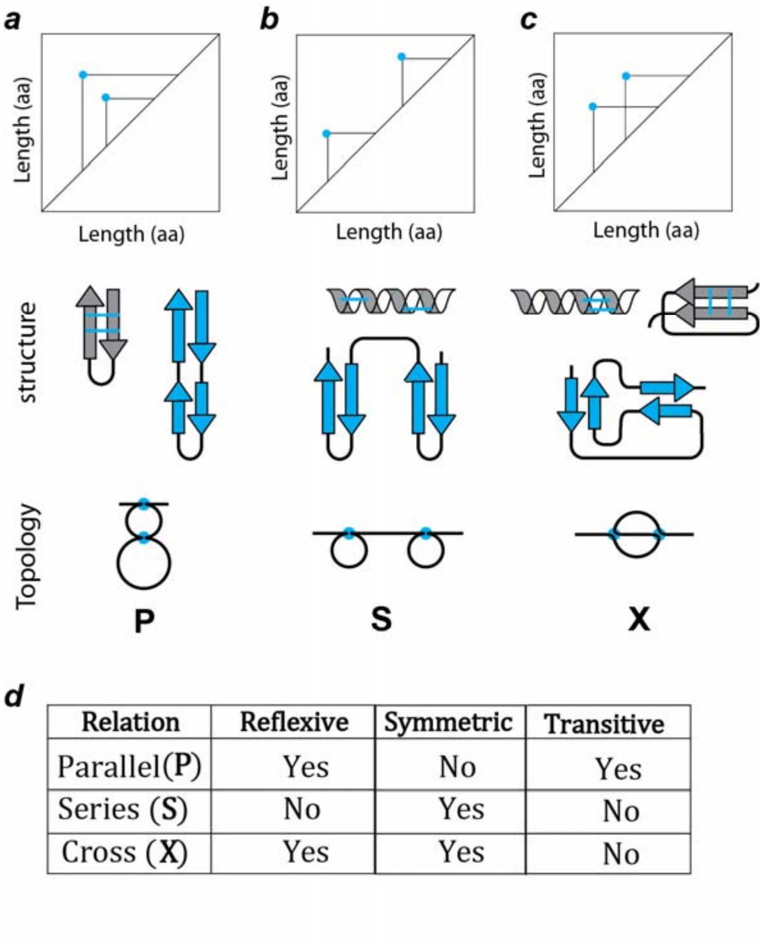
The notion of network topology(Goldenberg, 1999) has been used to quantify certain topological features of proteins. Here, one starts with identifying the residue-residue contacts within a protein structure. Because any two contacts are ultimately linked by the protein chain, the contacts can be seen as forming a network. These networks have been analyzed in terms of their statistical properties, such as node degree, clustering coefficients, betweenness, closeness centrality and contact order (Goldenberg, 1999). This approach has proven useful in predicting the folding rate of small proteins (Baker, 2000). However, a statistical quantification of topological features does not provide a description of the topology of a molecule as such, and does not address the issue of topological equivalence. Network topology ideas have been applied to distinguish small RNA structures (Melchers et al., 1997; Pasquali et al., 2005), but as we will show here this approach is less suited as a general framework to establish equivalence.

In this perspective paper, we propose a framework that allows one to formally describe the topology of folded linear chains, which we refer to as 'circuit topology' because of its conceptual similarities to topological features in electronic circuits. In the following sections we will describe its mathematical basis, how it resolves the recognition problem (Hass, 1998), and briefly discuss a number of possible applications.

2 The circuit topology approach

Folded biomolecular chains are typified by multiple intra-molecular contacts, in which one part of the chain binds to another part. The circuit topology notion that we propose is based on defining the pair-wise relations between these intra-molecular contacts. One can distinguish three types of relations between two contacts: parallel (P), series (S), and cross (X). For example, in a β-hairpin, any two contacts are arranged in parallel to each other, with one being contained in the loop created by the other (Fig. 2a). In contrast, in an α–helix, two distant contacts define two loops that are positioned in series with each other (Fig. 2b). When the two -helical contacts are close by, the loops they define may overlap, which gives rise to a cross arrangement (Fig. 2c). These contact arrangements are topologically distinct: any transition between them (for example from P to S) would involve the formation and rupture of contacts, while transitions within one topology can occur by continuous deformation.

The relations are general and can be applied to various systems and intra-molecular contacts. Central to the circuit topology approach is that one specifically chooses a type of contact, which is therefore by definition well-defined. For RNA structures, it is natural to define hybridized regions as contacts, while β-hairpins may be classified by contacts at the level of hydrogen bonds. In DNA origami, the anchoring points defined by the oligo 'staples' would be most relevant, while for chromosomes, it would rather be the contacts formed by associated proteins (e.g. CTCF protein(Richardson, 1977)). In metaloproteins such as zinc finger domains(Gamsjaeger et al., 2007) the metal-mediated contacts may be of central interest. In addition, the topologies of proteins may be classified by the arrangement of β-strand contacts (β-strand circuit topology; Fig. 2 a-c), disulfide bonds (disulfide bond circuit topology; Fig. S1), or contact regions with large interaction energies (Mashaghi et al., 2013).



Relation	Reflexive	Symmetric	Transitive
Parallel(**P**)	Yes	No	Yes
Series (**S**)	No	Yes	No
Cross (**X**)	Yes	Yes	No

Figure 2: Elementary topological relations between two contacts in a linear chain. Elementary relations are expressed by their contact map, illustrating structural motifs, and bubble graphs. (a) Parallel relation (P). (b) Series relation (S). (c) Cross relation (X). Choosing the type of contact is central. Different types of contacts give rise to different circuit topologies. Examples are shown for H-bonds and β-strand contacts. (d) Properties of topological relations. For example, S is symmetric because if contact A is in series with B, then B is also in series with A. However, the same does not hold for the parallel relation. Only P is transitive, which means that if A is in parallel with B and B with C, then A is also in parallel with C. According to the definitions (Box I), a contact is also in parallel with itself, and therefore the parallel relation is also reflexive. Contacts can not be in series with themselves, so the S relation is not reflexive. Note that none of the relations P, S and X satisfy all three properties. Interestingly however, P is both reflexive and transitive, and can thus be used to introduce the notion of "order" for contacts (Bloch, 2011) (Similarly ≤ is both reflexive and transitive and thus can order real numbers). These relation properties can be used to formulate rules that contacts within one folded linear chain must obey (see Box I).

The circuit topology relations (Fig. 2a-c) can be embedded within known set theory and discrete mathematics (see Box I). First, this allows one to rigorously define the topological relations between contacts in a chain. Second, it can be used to prove that the three relations form a complete set when considering binary contacts (see Box I). This means that they are necessary and sufficient to describe any folded structure with binary contacts. Note that one can readily extend this approach to also include contacts in which one site on the chain is bound to two other sites (Fig. S2, S3). Third, the circuit topology relations have specific properties, which inform on the rules that the contacts must obey. These rules are relevant to the recognition problem as well as engineering applications, as we will outline next.

3 Topology rules for folded chains

To illustrate how the properties of binary relations can be used, lets consider the Hammerhead ribozyme (Fig. 3a). For some pairs of contacts the relations are obvious when observing the structure. For instance, the yellow contact is in parallel with the green one, while green is in series with gray, as can be tabulated in a matrix (Fig. 3a, bottom). For other pairs of contacts the relations may be more difficult to establish, but can be inferred from known relations using their properties. The simplest example is the symmetric property of the series relation. Symmetry here means that if green is in series with gray, gray must also be in series with green. On the other hand, the parallel relation (P) is not symmetric: the loop formed by the yellow contact is enclosed in the loop formed by the green contact, but not *vice versa*. However, P is transitive. This means that if yellow is parallel to green, and green parallel to blue; then yellow is also parallel to blue (box I, chain rule I). Various rules are less intuitive, as seen in the following example. The red and gray contacts are crossing, and gray is in parallel with blue. This means that blue and red are either crossing or in parallel (Box I, chain rule V). But they cannot be in parallel: if they would be, the parallel nature of blue and gray would dictate red and gray must be also be parallel (Box I, chain rule I), which it isn't. Hence, blue and red must be crossing. Such algorithmic inference is useful in particular when structures become more complex, as for instance the pyramid-shaped protein origami that was recently realized (Fig. 3b) (Gradisar et al., 2013). Overall, these examples illustrate how complex recognition problems can be simplified by applying an algorithm of pre-defined rules on limited information.

The circuit topology rules can help engineering new molecules with specific topological features. For instance, within a loop of an RNA molecule A of known topology (the Hammerhead ribozyme), one may insert another molecule B (Fig. 3c). A priory it is not clear what the topology of the combined molecule AB is, in particular if the topology of B is unknown. However, one can use the topology rules to derive the topological relations. For instance, we can use the following information: yellow and gray in A are in series, and purple in B is in parallel with yellow in A, simply because B is inserted in the loop formed yellow. The rules then tell us that purple will be in series with gray (Box I, chain rule II). The other relations can be inferred in similar fashion. In general, the mathematical framework can help to systematically explore the space of possible topologies, design molecules with certain topologies, or modify the topologies of existing molecules.

The rules also provide information on topological constraints: the boundaries of topology space beyond which molecules cannot evolve or be engineered. Imagine chain A with two contacts in parallel, and chain B with two contacts in series. Inserting B into A gives rise to 4 additional contact relations. Regardless of how B is inserted, the rules show that these 4 relations may either be all P, all S, or 2 times S and 2 times P. However, having 1 time S with 3 times P is not possible because it would leads to a mathematical contradiction, where one must break the chain rules in Box I: chain rule II states that having 1 time S and 1 time P leads to a second S.

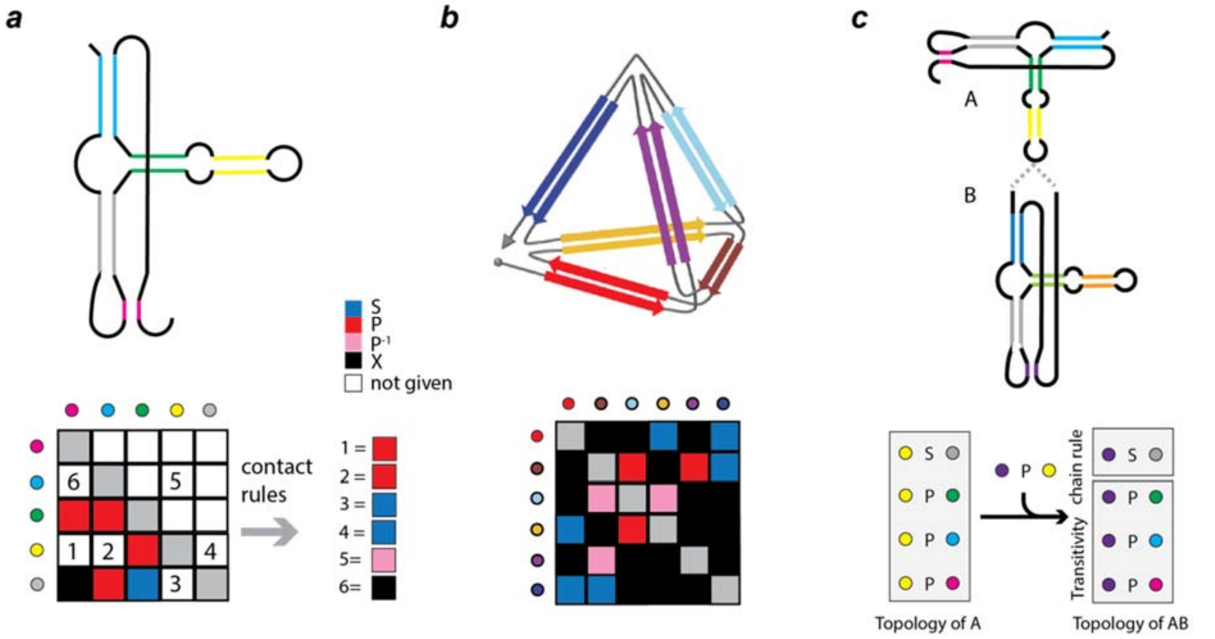


Figure 3: Determining the topology of folded linear chains (a) Structure of Hammerhead ribozyme. The topology matrix indicates how the row-contact relates to the column contact. Unknown relations can be derived using the topology rules in Fig 2 and Box I. aPb means that a is parallel to b. aP^{-1}b means that b is parallel to a. (b) Topology and DNA or protein origami. Topology of a polypeptide tetrahedron(Gradisar et al., 2013) is represented by its topology matrix. (c) Topological relation rules can help in engineering RNAs with arbitrary topologies. Here, an RNA chain with unknown topology (B) is inserted in the Hammerhead ribozyme (A). All contacts in B are parallel to the yellow contact in A. Because the relations between the yellow and all other contacts in A are known, we can deduce the topological relations between a contact in B (violet) with every contacts of A.

4 Topology and equivalence

A hallmark of a topology approach is whether it can determine whether two structures are topologically equivalent. Here we compare the resolving power of the circuit and network topologies in Fig. 4. The latter is described by a graph and an adjacency matrix (Gan et al., 2003), which indicates the number of direct connections between two contacts (Fig. 4). This network method correctly attributes different topologies to the PKB131 aptamer and a self-cleaving VS ribozyme. However, it also suggests that the latter is equivalent to a third 'test RNA', even though interconversion between the two clearly requires cutting and gluing of the RNA strand. In contrast, the circuit method correctly distinguishes the topologies of these two molecules, by presenting matrices that are not equivalent (Fig. 4).

Determining equivalence, or lack thereof, is central to understanding the huge diversity of biomolecular structures. Recent computational methods have made important progress in this issue by quantifying various geometric measures (Holm and Sander, 1996; Mizuguchi and Go, 1995), and hence to map the 'protein universe' of observed protein structures (Hou et al., 2005). It will be of interest to determine the relatedness of these proteins in terms of their circuit topology, and to perform comparisons with evolutionary, sequence, or functional relations. Take for example, the two proteins Xylanase 10c and Cellulase B (Fig. S3), which are different both in sequence and structure. However, they have identical beta-strand circuit topology (see Fig. S3). Interestingly, both bind carbohydrates and act as glycoside hydrolases. This example indicates that beta-strand circuit topology may be highly conserved and can inform on functionality. This could be consistent with the observation that within protein families, some features such as the length of peptide loops that are variable and do affect structure, do not affect beta-strand circuit topology. The circuit topology formalism allows one to rigorously compare the evolutionary conservation of topological versus non-topological features in proteins. Various mathematical methods exist to characterize and compare matrices, and hence these can be employed to quantify relatedness. We note that the circuit topology method is less suited to quantify relations between proteins that are very distant in terms of their circuit topology. For instance, when studying distances in terms of the beta-strand circuit topology, one cannot assess proteins that lack beta strands.

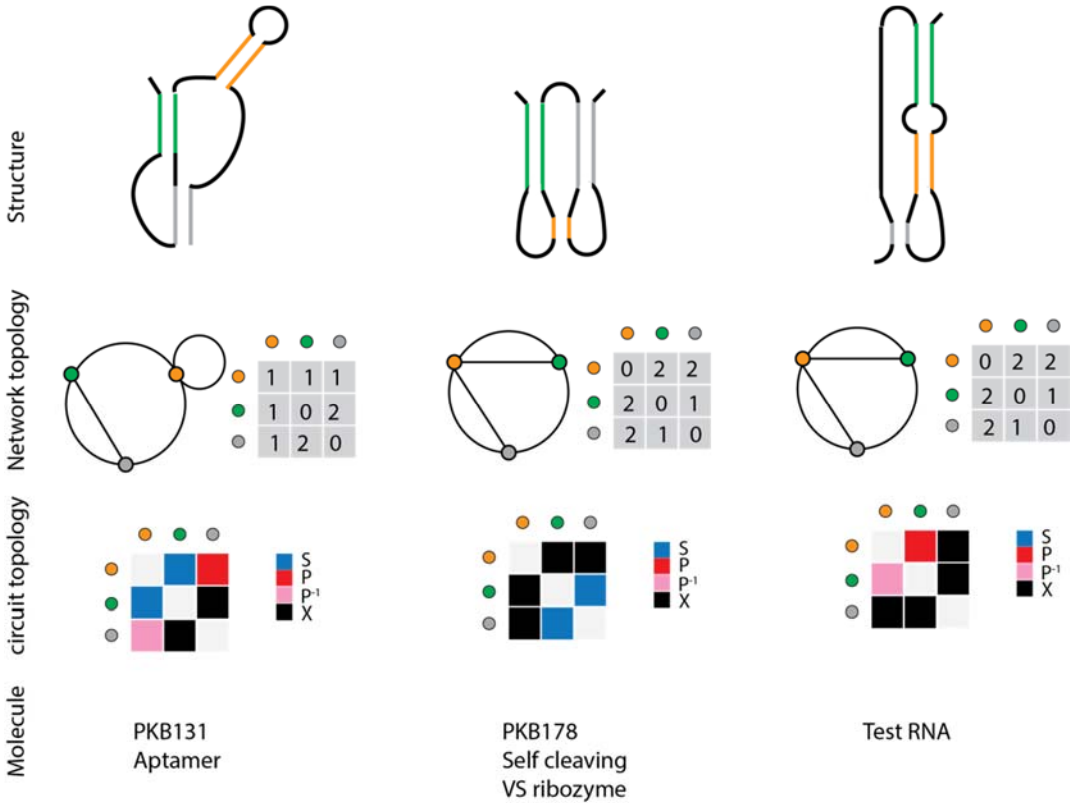


Figure 4: Topological equivalence. Determining whether two structures have an equivalent topology is a generic challenge, also for linear folded chains. Here we contrast the network topology (Gan et al., 2003) and the circuit topology approaches to distinguish folded structures. Network topology, with its dual graph and adjacency matrix indicating the number of direct contacts, properly distinguishes the first two structures, but erroneously suggests the latter two are equivalent. The circuit topology approach properly distinguishes all three structures.

-------- BOX I: Circuit topology relations --------

Definition of elementary circuit topology relations:

We consider a linear chain with contact sites numbered as $i = 1, 2, 3, \dots, n$. Contact C_1 connects sites *i* and *j*, and contact C_2 connects sites *r* and *s*. We define the following relations between C_1 and C_2:

Parallel: $C_1\mathbf{P}C_2 \Leftrightarrow [i,j] \subset [r,s]$

Series: $C_1\mathbf{S}C_2 \Leftrightarrow [i,j] \cap [r,s] = \varnothing$

Cross: $C_1\mathbf{X}C_2 \Leftrightarrow [i,j] \cap [r,s] \notin \{\varnothing, [i,j], [r,s]\}$

where $C_1\mathbf{P}C_2$ denotes C_1being parallel to C_2 (and similarly for **S** and **X**). The descriptions of the symbols are given in table 1.

Table 1. Set theory symbols

Set theory symbols	Description	Logical symbols	Description
$\subset$	Inclusion	$\forall$	For every
$\cap$	Intersection	$\wedge$	And
$[i,j]$	An interval of natural numbers from *i* to *j*	$\vee$	Or
$\varnothing$	Empty set	$\sim$	Not
$\in$	Belongs	$\Longrightarrow$	Implies (if then)
$\mathbb{N}$	Set of natural numbers	$\Leftrightarrow$	If and only if

Proof of completeness:

To demonstrate that S, P, and X are sufficient and necessary to describe the topology of any folded linear chain with binary contacts, we aim to show that if two contacts are not in parallel, then they must be either in series or cross:

$$C_1{\sim}\mathbf{P}C_2 \wedge C_2{\sim}\mathbf{P}C_1 \Longrightarrow [i,j] \not\subset [r,s] \wedge [r,s] \not\subset [i,j] \Longrightarrow [i,j] \cap [r,s] \notin \{[i,j], [r,s]\}$$

$$[i,j] \cap [r,s] = \varnothing \Longrightarrow \quad C_1\mathbf{S}C_2$$

$$[i,j] \cap [r,s] \neq \varnothing \Longrightarrow \quad C_1\mathbf{X}C_2$$

Topology rules:

For any arbitrary choice of contacts the following rules apply.

Chain rule I: $[C_1\mathbf{P}C_2 \wedge C_2\mathbf{P}C_3] \Longrightarrow C_1\mathbf{P}C_3$

which can be generalized to: $[C_1\mathbf{P}C_2 \wedge C_2\mathbf{P}C_3 \wedge \dots \wedge C_v\mathbf{P}C_{v+1}] \Longrightarrow C_1\mathbf{P}C_{v+1}$. Note that **S** and **X** relations are not transitive.

Chain rule II: $[C_1\mathbf{P}C_2 \wedge C_2\mathbf{S}C_3] \Longrightarrow C_1\mathbf{S}C_3$

Chain rule III: $[C_1\mathbf{P}C_2 \wedge C_2\mathbf{X}C_3] \Longrightarrow C_3{\sim}\mathbf{P}C_1 \wedge C_1{\sim}\mathbf{P}C_3$ (i.e. $C_1\mathbf{S}C_3 \vee C_1\mathbf{X}C_3$)

Chain rule IV: $[C_1\mathbf{S}C_2 \wedge C_2\mathbf{X}C_3] \Longrightarrow C_3{\sim}\mathbf{P}C_1$ (i.e. $C_1\mathbf{S}C_3 \vee C_1\mathbf{X}C_3 \vee C_1\mathbf{P}C_3$)

Chain rule V: $[C_1\mathbf{X}C_2 \wedge C_2\mathbf{P}C_3] \Longrightarrow C_1{\sim}\mathbf{S}C_3 \vee C_3{\sim}\mathbf{P}C_1$ (i.e. $C_1\mathbf{X}C_3 \vee C_1\mathbf{P}C_3$)

Chain rule VI: $[C_1\mathbf{S}C_2 \wedge C_2\mathbf{P}C_3] \Longrightarrow C_3{\sim}\mathbf{P}C_1$ (i.e. $C_1\mathbf{P}C_3 \vee C_1\mathbf{S}C_3 \vee C_1\mathbf{X}C_3$)

-------- End of BOX I --------

5 Allowed and forbidden transitions

A generic mechanism to recognize the presence of knots within a chain is to pull at the two termini (Taylor and Lin, 2003). Here we explore whether pulling can also be used to distinguish topologies of folded chains. Pulling at the termini in this case leads to a successive rupturing of contacts, during which one can for instance monitor the length between termini.

As an illustration, one can consider pulling on P, S, and X structures (Fig. 5a). P can only be broken down *via* one specific length, because one loop is enclosed in the other and thus will not rupture nor yield a length change. In contrast, for S two pathways are possible because the two loops are independent. On the other hand, the two corresponding intermediate states cannot interconvert, as this would involve not only cutting but also gluing contacts. The cross-topology presents yet another cause for order: a contact may experience tension only in one direction and hence will not rupture like the contacts that experience tension in two opposite directions. In analogy with spectroscopy, one can refer to these constraints as allowed and forbidden transitions, and tabulate them in a transition matrix. The matrices are indeed specific to the topology (Fig. 5b). This differentiation does not rely on length but rather on topology information: Even if we shrink and stretch every segment to make the lengths identical, the topologies can be resolved by pulling (Fig. 5c). Consequently, molecules with different length and sequence but identical topology can be assigned correctly into the same topological class.

For a chain described by the binary topology relations (P, S, X), the number of possible unfolding paths can be calculated analytically. A parallel relation only allows one route while series and cross relations allow for two routes each. The number of possible pathways will therefore be 2 to the power of the total number of S and X relations (see captions Fig. 5c). In folded biomolecules, one site may be involved in two or more contacts. For instance, when considering circuit topologies defined by contacts between beta strands, one strand within a beta sheet contacts two other strands. Such two contacts may display cooperativity: breaking one will influence the stability of the remaining contact. When two contacts that share a contact site are in series, we describe it as a concerted series relation (see also Fig. S2). When two contacts that share a contact site are in parallel, we describe it as a concerted parallel relation. Note that two contacts that share a contact site cannot be in a cross relation. In molecules with such concerted relations, the number of pathways is then different, but can still be calculated.

We have used pulling so far as a mathematical operation. However, the same notions can be applied to understand length transitions in the mechanical unfolding of proteins and RNA structures by single-molecule methods (Liphardt et al., 2001; Stigler et al., 2011). Different contacts will then exhibit different contact free energies, and those with lower energies will break earlier. While topology sets the selection rules, energy thus affects transition probabilities. In single-molecule mechanical unfolding assays involving large proteins it is often a challenge to relate the many observed length transitions to structure. The tools presented here can provide hypothesis about the order in which contacts are disrupted by pulling such molecules.

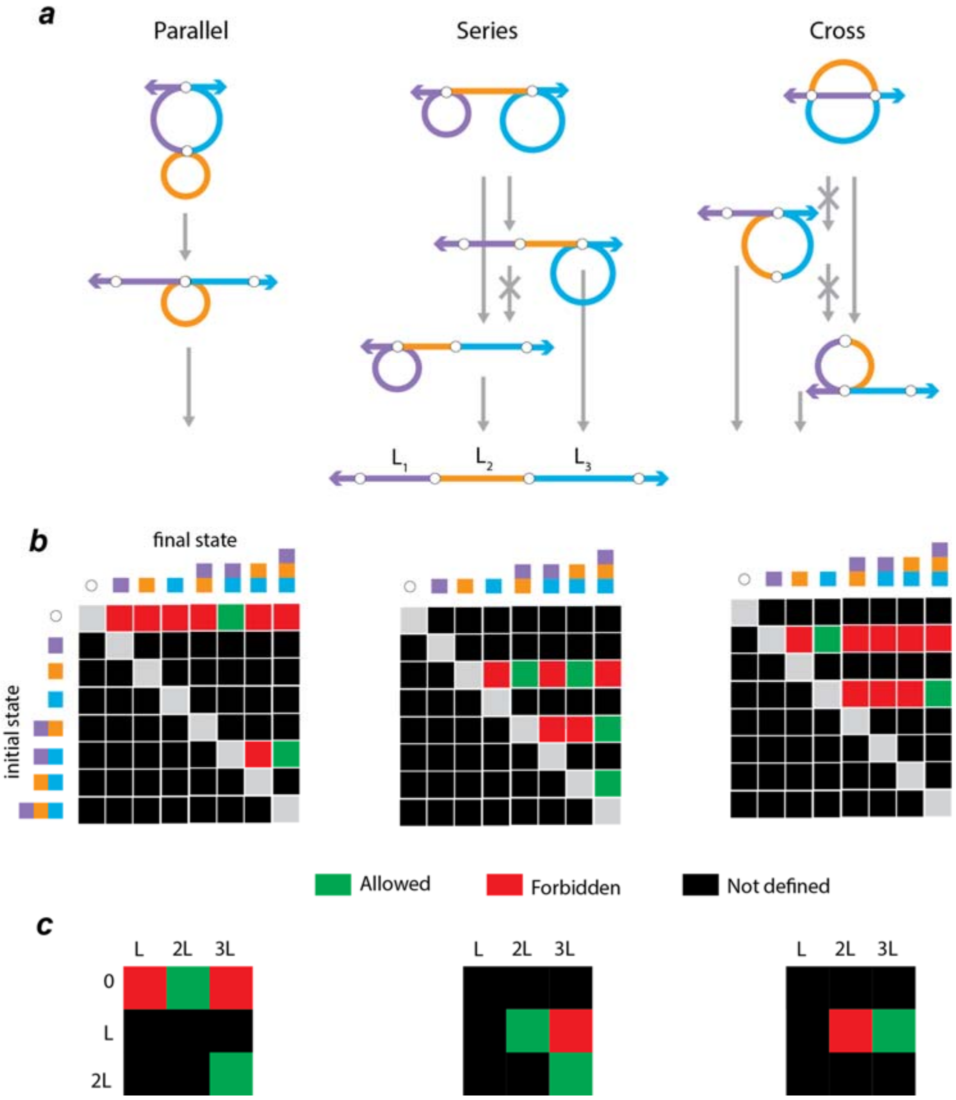


Figure 5: Allowed and forbidden transitions. (a) Successive states when rupturing parallel, series, and cross topologies by pulling at the termini. Some transitions between states will not occur (crossed-out arrows) because they involve not only cutting but also gluing (series, second crossed arrow in cross), or because they are not promoted by pulling (first crossed arrow in cross). (b) Matrices indicating transitions between two states for $L_1<L_2<L_3$. Without losing generality we assume that $L_3<L_1+L_2$. States are denoted by the color of the segments that contribute to the length of the stretched chain. (c) Transition matrices for equal L_i. In both cases, the matrices are specific to the topology and can thus be used to differentiate. The number of unfolding pathways in a fold with binary contacts is given by $1^{Np}2^{(Nx + Ns)}$.

6 Topology and folding

Topology can have diverse functional consequences. For instance, one can consider the constraints it imposes on the conformational search during folding. Which generic properties determine folding rates of proteins is still a highly debated issue. However there is an agreement that the folding rate correlates with the properties of native fold (Faisca et al., 2012). It has been shown that the folding rates strongly correlate with contact order for small proteins (Baker, 2000) while, for large proteins, size (length) is the best found determinant of the rate (Ivankov et al., 2003). Circuit topology of the natively folded molecule could be a property that affects the folding rate (Fig. 6). A simple example is that contacts that are in series can develop independently, while contacts in parallel will have a tendency to form the most nested contacts first. In addition to protein folding, evidences exist for the role of topology in RNA folding. In tRNA, formation of anti-codon, T- and D-stems often facilitate formation of acceptor stem to which they are parallel. However they do not facilitate formation of each other, as they are in series (Richardson, 1981).

Next, we explore the different roles of contact order and circuit topology in folding. We consider equal-size idealized chains with two contacts of the same binding affinity (Fig. 6). The distance between contact sites sets the time for the contact to form: closely spaced contact sites find each other rapidly, while distant sites are slow to form contacts. Interestingly, one can arrange these contacts such that the contact order is identical, but the folding rates are different (Fig. 6, chain A and B). Moreover, the contact order may be higher (than chain A) while the folding time is lower (Fig. 6, chain C), which is opposite to the reported dependence. These elementary examples suggest that the circuit topology contains information on folding rates that is not captured in the contact order (Baker, 2000).

Circuit topology does not necessarily inform on the free energy. The formation of a loop in an idealized chain comes with and entropic cost that scales with the inverse of the loop size (Muthukumar, 1999; Pauling and Corey, 1951a, b). Because changes in loop size do not affect topology, one can deform chain B to increase one of the loops and obtain chain C, while keeping the topology and size unaffected. Folding is obviously determined by numerous other factors. Here we merely propose that -like contact order- circuit topology is a generic property that imposes quantifiable constraints on the folding and unfolding of proteins and other biomolecular systems. The notions discussed here can be generalized to less intuitive cases, where the topologies are composed of multiple parallel, series, and cross motifs, in order to explore the role of topology in folding as well as misfolding (Lodge and Muthukumar, 1996).

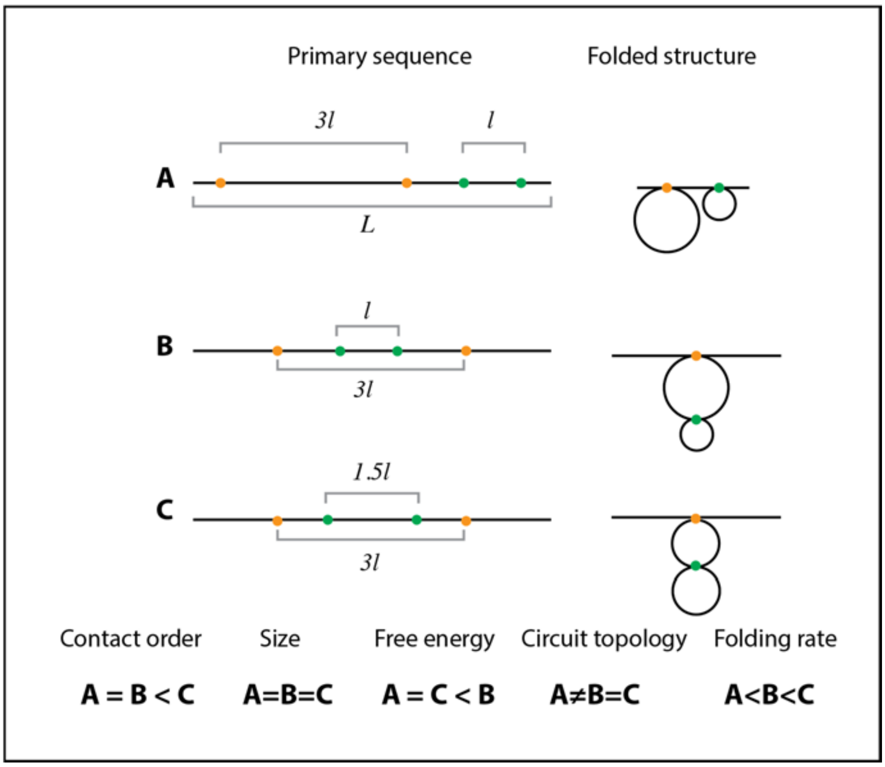


Figure 6: Circuit topology is a determinant of folding rate. Idealized chains A and B form two contacts, orange and green. Both chains have identical contact orders and size but distinct topologies and folding rates. For a contact to form, the contact sites have to search in 3D space to find each other, a process that depends on their distance in the non-looped primary sequence ($\tau \sim l^{3/2}$ for a freely jointed chain (Szabo et al., 1980). Here we ignore the tail effect (Doucet et al., 2007; Fierz and Kiefhaber, 2007) for simplicity.). When all contacts are formed the chain will be in its natively folded state. The folding speed is dictated by one or more properties of the native fold: i.e. size, contact order, free energy (entropy) and topology. Circuit topology is indeed a determinant of the folding rate. To demonstrate this, we compare the folding rates of A with B and C. The fact that B and C fold faster than A, cannot be explained by the size, contact order and entropy of their native conformations. Thus topology determines the folding rate difference. When two chains are topologically equivalent, the folding rates are determined by other determinants (e.g. when B and C are compared). It can be shown that the chains with parallel topologies fold faster than the one with series topology. Chain C has identical free energy (entropy) as chain A but different topologies, and different contact order. Surprisingly the chain that has smaller contact order folds slower.

7 Molecular complexes and chromosome structure

The challenge to understand architecture and diversity of folded biomolecules extend beyond the individual molecule to molecular complexes. For instance, the ubiquitously expressed essential cofactor ldb1, which plays diverse roles in development, binds many proteins with LIM-domain in a specific fashion. By binding in extended fashion, it extends repeated β-sheets in its binding partner (Fig. 7a). This β-zipper motif is also observed in fibronectin-binding peptides, which bind fibronectin in a similar fashion (Ryan and Matthews, 2005). Another generic molecular interaction is the RNA kissing complex. This interaction shows high stability (Li et al., 2006) and is seen in functions ranging from viral genome replication (Chang and Tinoco, 1994) to RNA synthesis (Melchers et al., 1997). In these structures, some essential minimal structural features can be distinguished. In β-zippers the peptide ligand adds a β-strand to an existing array of beta sheets of its binding partner, and in kissing complexes binding occurs between two loops formed by RNA hairpins. However, it is not straightforward to formally define these generic features. Doing so would enable a more systematic and precise analysis of structural similarity, and its relation to biological function.

Although we do not develop this rigorously here, the notion of circuit topology could be extended to describe complexes involving more than one molecule (Fig. 7). For instance, one can imagine the ends of the two molecules to be physically linked, which allows one to consider the topological relations within this larger single molecule. Such an approach could provide a new perspective to molecular interactions that alter the topology of a molecule. Examples are proteins that are intrinsically disordered protein in isolation but adopt a folded state upon interaction with a binding partner, or the binding of a coordinating metal ion that adds a contact and hence changes the topology. Chaperone-guided folding is another phenomenon where interactions with a second molecule affect their folding states. How chaperones affect the conformational state during folding is now starting to be addressed, for instance using single-molecule approaches(Mashaghi et al., 2013). Using these novel experimental techniques and topology notions, one can begin to address whether chaperones assist folding by transiently modulating the topology of their clients.

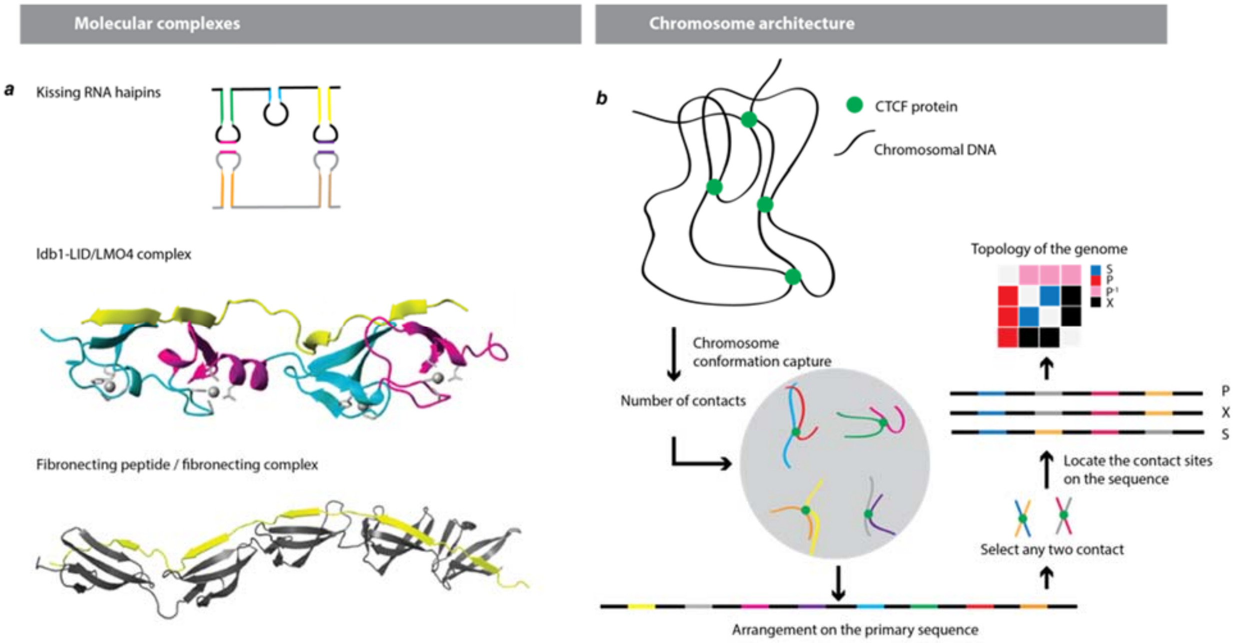


Figure 7: Circuit topology of molecular complexes. (a) Arrangement of intra and inter molecular contacts in RNA-RNA and protein-protein complexes. An interacting molecule may modulate the topology of its target, by mediating contacts between two (distant) residues. Binding may also change the stability of an existing intra molecular contact within the client. The tandem β-zipper Idb1 (unstructured in isolation) binds specifically the β-β contacts of its partner the LIM-only protein (LMO4, PDB: 1RUT) (Daw, 2013). The LIM-domain-binding protein Idb1 is a ubiquitously expressed essential cofactor that plays important roles in the development of complex organisms (Matthews and Visvader, 2003). β-zipper motif is also observed in fibronectin-binding peptides. (b) Possible approach to analyze the topology of chromosomes using chromosome confirmation capture (van Steensel and Dekker, 2010).

Understanding the architecture of chromosomes has recently become an area of intense research (van Steensel and Dekker, 2010). Chromosome architecture must faithfully be re-established every cell cycle, and is increasingly being implicated in human pathologies (Engreitz et al., 2012; Zhang et al., 2012). New data is emerging from innovative technologies such as fluorescence in situ hybridization (FISH), in vivo tagging of genomic loci and 3C-based technologies. They are underscoring that protein-mediated linking of distant chromosomal loci plays an important role, which suggests that chromosomes exhibit specific topologies (Fig. 7b). Circuit topology and its tools (Fig. 2-5) could provide a powerful tool to interpret these data, analyze equivalence of chromosome architectures for different conditions. 3C-based technologies in particular are well suited to analyze chromosomal topologies, as the coincidence of sequences that they provide can be mapped onto the known chromosome sequence, and hence can be used directly to determine the circuit topology matrix (see Fig. 7b).

8 Conclusions

Chemists early on recognized implications of shapes in the chemistry of proteins and macromolecules. In his 1974 Nature article on molecular basis of biological specificity, Linus Pauling wrote (Pauling, 1974): "I am convinced that it will be found in the future, [...] that the shapes and sizes of molecules are of just as great significance in determining their physiological behavior as are their internal structure and ordinary chemical properties. I believe that the thorough investigation of the shapes and sizes of molecules will lead to great advances in fundamental biology and medicine." Now after four decades, the notion of topology has already emerged as a powerful concept to describe the essence of complex molecular structures and guided synthesis of materials with interesting properties(Kamien, 2003; Siegel, 2004; Siegel et al., 2012). Here we have aimed to briefly review some of the existing approaches in molecular topology, and to introduce an extension that provides an integral description of the topology of folded linear chains and allows analysis of equivalence. The extension allows the difficult problem of self-interacting chain topology to be rigorously addressed. This general framework can be applied in addressing a wide range of molecular systems and scientific questions.

Acknowledgments

This work 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

Ayme, J.F., Beves, J.E., Leigh, D.A., McBurney, R.T., Rissanen, K., and Schultz, D. (2012). A synthetic molecular pentafoil knot. Nature chemistry *4*, 15-20.

Bailor, M.H., Sun, X., and Al-Hashimi, H.M. (2010). Topology links RNA secondary structure with global conformation, dynamics, and adaptation. Science *327*, 202-206.

Baker, D. (2000). A surprising simplicity to protein folding. Nature *405*, 39-42.

Blankenship, J.W., and Dawson, P.E. (2007). Threading a peptide through a peptide: protein loops, rotaxanes, and knots. Protein science : a publication of the Protein Society *16*, 1249-1256.

Bloch, E.D. (2011). Proofs and fundamentals : a first course in abstract mathematics, 2nd edn (New York: Springer).

Brown, I.D. (2002). Topology and chemistry. Struct Chem *13*, 339-355.

Cavalli, G., and Misteli, T. (2013). Functional implications of genome topology. Nature structural & molecular biology *20*, 290-299.

Chang, K.Y., and Tinoco, I. (1994). Characterization of a Kissing Hairpin Complex Derived from the Human-Immunodeficiency-Virus Genome. P Natl Acad Sci USA *91*, 8705-8709.

Coskun, A., Banaszak, M., Astumian, R.D., Stoddart, J.F., and Grzybowski, B.A. (2012). Great expectations: can artificial molecular machines deliver on their promise? Chemical Society reviews *41*, 19-30.

Daw, R. (2013). Materials science: Topology matters. Nature *493*, 168.

Doucet, D., Roitberg, A., and Hagen, S.J. (2007). Kinetics of internal-loop formation in polypeptide chains: A simulation study. Biophys J *92*, 2281-2289.

Engreitz, J.M., Agarwala, V., and Mirny, L.A. (2012). Three-Dimensional Genome Architecture Influences Partner Selection for Chromosomal Translocations in Human Disease. PloS one *7*.

Faisca, P.F.N., Travasso, R.D.M., Parisi, A., and Rey, A. (2012). Why Do Protein Folding Rates Correlate with Metrics of Native Topology? PloS one *7*.

Fierz, B., and Kiefhaber, T. (2007). End-to-end vs interior loop formation kinetics in unfolded polypeptide chains. J Am Chem Soc *129*, 672-679.

Flapan, E. (2000). When topology meets chemistry : a topological look at molecular chirality (Cambridge ; New York Washington, DC: Cambridge University Press ; Mathematical Association of America).

Francl, M. (2009). Stretching topology. Nature chemistry *1*, 334-335.

Gamsjaeger, R., Liew, C.K., Loughlin, F.E., Crossley, M., and Mackay, J.P. (2007). Sticky fingers: zinc-fingers as protein-recognition motifs. Trends Biochem Sci *32*, 63-70.

Gan, H.H., Pasquali, S., and Schlick, T. (2003). Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. Nucleic Acids Res *31*, 2926-2943.

Goldenberg, D.P. (1999). Finding the right fold. Nat Struct Biol *6*, 987-990.

Gradisar, H., Bozic, S., Doles, T., Vengust, D., Hafner-Bratkovic, I., Mertelj, A., Webb, B., Sali, A., Klavzar, S., and Jerala, R. (2013). Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. Nature chemical biology.

Guan, Z.B., Cotts, P.M., McCord, E.F., and McLain, S.J. (1999). Chain walking: A new strategy to control polymer topology. Science *283*, 2059-2062.

Harada, A. (2012). Supramolecular polymer chemistry (Weinheim, Germany: Wiley-VCH).

Hass, J. (1998). Algorithms for recognizing knots and 3-manifolds. Chaos Soliton Fract *9*, 569-581.

Holm, L., and Sander, C. (1996). Mapping the protein universe. Science *273*, 595-602.

Hou, J.T., Jun, S.R., Zhang, C., and Kim, S.H. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. P Natl Acad Sci USA *102*, 3651-3656.

Hutchinson, E.G., and Thornton, J.M. (1993). The Greek Key Motif - Extraction, Classification and Analysis. Protein Eng *6*, 233-245.

Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D., and Finkelstein, A.V. (2003). Contact order revisited: Influence of protein size on the folding rate. Protein Science *12*, 2057-2062.

Kamien, R.D. (2003). Topology from the bottom up. Science *299*, 1671-1673.

Kauffman, L.H. (1994). Tales of topology. Science *265*, 2108-2110.

Li, P.T.X., Bustamante, C., and Tinoco, I. (2006). Unusual mechanical stability of a minimal RNA kissing complex. P Natl Acad Sci USA *103*, 15847-15852.

Liang, C., and Mislow, K. (1995). Topological features of human chorionic gonadotropin. Biopolymers *35*, 343-345.

Liphardt, J., Onoa, B., Smith, S.B., Tinoco, I., Jr., and Bustamante, C. (2001). Reversible unfolding of single RNA molecules by mechanical force. Science *292*, 733-737.

Lodge, T.P., and Muthukumar, M. (1996). Physical chemistry of polymers: Entropy, interactions, and dynamics. J Phys Chem-Us *100*, 13275-13292.

Lopez, V., Martinez-Robles, M.L., Hernandez, P., Krimer, D.B., and Schvartzman, J.B. (2012). Topo IV is the topoisomerase that knots and unknots sister duplexes during DNA replication. Nucleic Acids Res *40*, 3563-3573.

MacBeath, G., Kast, P., and Hilvert, D. (1998). Redesigning enzyme topology by directed evolution. Science *279*, 1958-1961.

Mallam, A.L., Rogers, J.M., and Jackson, S.E. (2010). Experimental detection of knotted conformations in denatured proteins. P Natl Acad Sci USA *107*, 8189-8194.

Manoil, C., and Beckwith, J. (1986). A genetic approach to analyzing membrane protein topology. Science *233*, 1403-1408.

Mashaghi, A., Kramer, G., Bechtluft, P., Zachmann-Brand, B., Driessen, A.J.M., Bukau, B., and Tans, S.J. (2013). Reshaping of the conformational search of a protein by the chaperone trigger factor. Nature.

Matthews, J.M., and Visvader, J.E. (2003). LIM-domain-binding protein 1: a multifunctional cofactor that interacts with diverse proteins. Embo Rep *4*, 1132-1137.

Melchers, W.J.G., Hoenderop, J.G.J., Slot, H.J.B., Pleij, C.W.A., Pilipenko, E.V., Agol, V.I., and Galama, J.M.D. (1997). Kissing of the two predominant hairpin loops in the coxsackie B virus 3' untranslated region is the essential structural feature of the origin of replication required for negative-strand RNA synthesis. J Virol *71*, 686-696.

Meyer, C.D. (2000). Matrix analysis and applied linear algebra (Philadelphia: Society for Industrial and Applied Mathematics).

Milner-White, E.J., and Poet, R. (1986). Four classes of beta-hairpins in proteins. The Biochemical journal *240*, 289-292.

Mizuguchi, K., and Go, N. (1995). Seeking significance in three-dimensional protein structure comparisons. Curr Opin Struct Biol *5*, 377-382.

Moutevelis, E., and Woolfson, D.N. (2009). A Periodic Table of Coiled-Coil Protein Structures. J Mol Biol *385*, 726-732.

Muthukumar, M. (1999). Chain entropy: Spoiler or benefactor in pattern recognition? P Natl Acad Sci USA *96*, 11690-11692.

Noel, J.K., Onuchic, J.N., and Sulkowska, J.I. (2013). Knotting a Protein in Explicit Solvent. The Journal of Physical Chemistry Letters *4*, 3570-3573.

Onuchic, J.N., and Wolynes, P.G. (2004). Theory of protein folding. Curr Opin Struc Biol *14*, 70-75.

Pasquali, S., Gan, H.H., and Schlick, T. (2005). Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. Nucleic Acids Res *33*, 1384-1398.

Pauling, L. (1974). Molecular-Basis of Biological Specificity. Nature *248*, 769-771.

Pauling, L., and Corey, R.B. (1951a). Configurations of Polypeptide Chains with Favored Orientations around Single Bonds - 2 New Pleated Sheets. P Natl Acad Sci USA *37*, 729-740.

Pauling, L., and Corey, R.B. (1951b). The Pleated Sheet, a New Layer Configuration of Polypeptide Chains. P Natl Acad Sci USA *37*, 251-256.

Ponnuswamy, N., Cougnon, F.B.L., Clough, J.M., Pantos, G.D., and Sanders, J.K.M. (2012). Discovery of an Organic Trefoil Knot. Science *338*, 783-785.

Rapp, M., Granseth, E., Seppala, S., and von Heijne, G. (2006). Identification and evolution of dual-topology membrane proteins. Nature structural & molecular biology *13*, 112-116.

Richardson, J.S. (1977). beta-Sheet topology and the relatedness of proteins. Nature *268*, 495-500.

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. Advances in protein chemistry *34*, 167-339.

Rosa, A. (2012). Topological jamming of spontaneously knotted polyelectrolyte chains driven through a nanopore. Physical Review Letters.

Ryan, D.P., and Matthews, J.M. (2005). Protein-protein interactions in human disease. Curr Opin Struct Biol *15*, 441-446.

Sabato, J.A. (1970). Topology and metallurgy. Nature *227*, 757.

Sali, A., Shakhnovich, E., and Karplus, M. (1994). How Does a Protein Fold. Nature *369*, 248-251.

Senyuk, B., Liu, Q., He, S., Kamien, R.D., Kusner, R.B., Lubensky, T.C., and Smalyukh, II. (2013). Topological colloids. Nature *493*, 200-205.

Shank, E.A., Cecconi, C., Dill, J.W., Marqusee, S., and Bustamante, C. (2010). The folding cooperativity of a protein is controlled by its chain topology. Nature *465*, 637-U134.

Shortle, D., and Ackerman, M.S. (2001). Persistence of native-like topology in a denatured protein in 8 M urea. Science *293*, 487-489.

Sibanda, B.L., and Thornton, J.M. (1991). Conformation of Beta-Hairpins in Protein Structures - Classification and Diversity in Homologous Structures. Method Enzymol *202*, 59-82.

Siegel, J.S. (2004). Chemical topology and interlocking molecules. Science *304*, 1256-1258.

Siegel, J.S., Liang, C.Z., Mislow, K., and Am, J. (2012). Driving the formation of molecular knots (vol 338, pg 752, 2012). Science *338*, 1287-1287.

Skrbic, T., Micheletti, C., and Faccioli, P. (2012). The role of non-native interactions in the folding of knotted proteins. Plos Comput Biol *8*, e1002504.

Soler, M.A., and Faisca, P.F. (2013). Effects of knots on protein folding properties. PloS one *8*, e74755.

Stigler, J., Ziegler, F., Gieseke, A., Gebhardt, J.C.M., and Rief, M. (2011). The Complex Folding Network of Single Calmodulin Molecules. Science *334*, 512-516.

Sulkowska, J.I., Rawdon, E.J., Millett, K.C., Onuchic, J.N., and Stasiak, A. (2012). Conservation of complex knotting and slipknotting patterns in proteins. Proc Natl Acad Sci U S A *109*, E1715-1723.

Sulkowska, J.I., Sulkowski, P., Szymczak, P., and Cieplak, M. (2008). Stabilizing effect of knots on proteins. Proc Natl Acad Sci U S A *105*, 19714-19719.

Sumners, D.W., and Whittington, S.G. (1988). Knots in Self-Avoiding Walks. J Phys a-Math Gen *21*, 1689-1694.

Szabo, A., Schulten, K., and Schulten, Z. (1980). First passage time approach to diffusion controlled reactions. Journal of Chemical Physics *72*, 4350

Taylor, W.R. (2000). A deeply knotted protein structure and how it might fold. Nature *406*, 916-919.

Taylor, W.R. (2002). A 'periodic table' for protein structures. Nature *416*, 657-660.

Taylor, W.R., and Lin, K. (2003). Protein knots - A tangled problem. Nature *421*, 25-25.

Terentjev, E. (2013). Liquid crystals: Interplay of topologies. Nature materials *12*, 187-189.

Tezuka, Y., and Oike, H. (2001). Topological polymer chemistry: Systematic classification of nonlinear polymer topologies. J Am Chem Soc *123*, 11570-11576.

van Steensel, B., and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. Nat Biotechnol *28*, 1089-1095.

VanLoock, M.S., Harris, B.A., and Harvey, S.C. (1998). To knot or not to knot? Examination of 16S ribosomal RNA models. J Biomol Struct Dyn *16*, 709-713.

Virnau, P., Mirny, L.A., and Kardar, M. (2006). Intricate knots in proteins: Function and evolution. Plos Comput Biol *2*, 1074-1079.

von Heijne, G. (2006). Membrane-protein topology. Nature reviews. Molecular cell biology *7*, 909-918.

Wetlaufe.Db. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. P Natl Acad Sci USA *70*, 697-701.

Yamamoto, T. (2012). Synthesis of cyclic polymers and topology effects on their diffusion and thermal properties. Polymer Journal 1-7.

Yan, H., Zhang, X., Shen, Z., and Seeman, N.C. (2002). A robust DNA mechanical device controlled by hybridization topology. Nature *415*, 62-65.

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations. Cell *148*, 908-921.