Topology of folded molecular chains: from single biomolecules to engineered origami

Scalvini, B.; Sheikhhassani, V.; Woodard, J.; Aupič, J.; Dame, R.T,; Jerala, R.; Mashaghi, A.

CellPress
REVIEWS

Opinion

Topology of Folded Molecular Chains: From Single Biomolecules to Engineered Origami

Barbara Scalvini,[1] Vahid Sheikhhassani,[1] Jaie Woodard,[1] Jana Aupič,[2]
Remus T. Dame,[3] Roman Jerala,[2] and Alireza Mashaghi[1,*]

The topology of biological polymers such as proteins and nucleic acids is an important aspect of their 3D structure. Recently, two applications of topology to molecular chains have emerged as important theoretical developments that are beginning to find utility in heteropolymer characterization and design: namely, circuit topology (CT) and knot theory. Here, we review the application of these two theories to protein, RNA, and DNA/genome structure, focusing on connections to conventional 3D structural information and relevance to function and highlighting recent experimental findings. We conclude with a discussion of recent applications to molecular origami and engineering.

Highlights

Circuit topology and knot theory are mathematically rigorous ways of describing the topology of a folded molecular chain. Conversions between topological states can be understood in terms of simple rules within developed mathematical frameworks.

The circuit topology of proteins and changes to their topology can be readily extracted from Protein Data Bank structures. The circuit topology of proteins underlies their evolution, folding, functionally relevant structures, and dynamics.

Knotted proteins exhibit distinct cellular, thermodynamic, and kinetic properties and are evolutionarily conserved. Studies of knotted polymers yield information about folding and molecular structure more generally.

Protein origami design principles were defined and provided in the form of a computational platform for the design of arbitrary complex CCPO polyhedra.

Topology: A Key Property to Disentangle Folding Complexity

Despite their apparent simplicity, linear heteropolymer chains may fold into distinct topologically diverse structures. In polymer chemistry, the diverse collection of linear polymers is supplemented by branched and cyclical structures, while in biological chemistry linear protein and nucleic acid chains adopt various topologies via chain folding. Folding involves rearrangements of the chain and the formation of contacts. In biology, we encounter a vast multiplicity of folded polymer identities, with chemical and structural, as well as functional, relations. Topology can not only help us make sense of the complex network of structural relationships, but can provide insights into folding mechanisms, conformational dynamics, and folding stability, ultimately aiding protein and drug design [1–4]. A particular challenge, and an opportunity for innovation, has been the application of topology to folded linear chains where 3D structures are stabilized by noncovalent intrachain contacts [5].

In this Opinion article, we highlight two recent applications of topology in the biomolecular sciences and molecular engineering. These topological approaches promise to categorize linear polymer structures. Knot theory categorizes molecular structures based on whether and how they are knotted. CT, in the context of polymer structure, categorizes folded linear chains based on their contact arrangement (Figure 1A), allowing structural summaries and comparisons in terms of topological building blocks and sets of permutation operations.

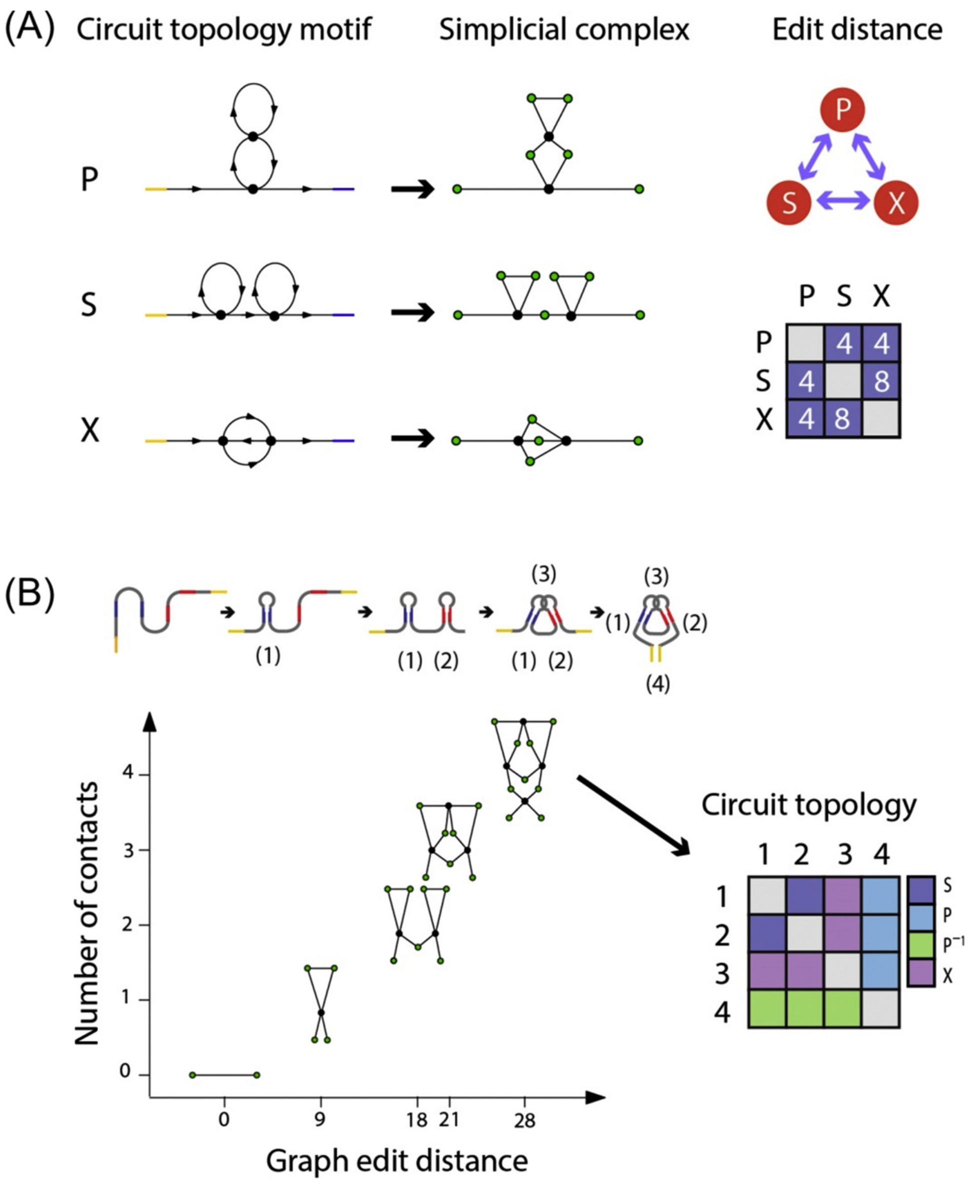
Knot Theory and CT: Basic Definitions

Formally, a knot is an embedding of the circle in 3D space. A knot may be equivalent (through stretching and bending operations, without allowing the knot to pass through itself) to the trivial knot, or circle, or to other knots with greater minimal numbers of crossings in their projections onto the plane. In contrast to proteins, RNA, and linear DNA, such knots lack a start and end point. However, linear molecules, on connecting the endpoints across an external arc traversing the 3D surface, may be said to be knotted or unknotted (trivial knot), according to the topology of the backbone [6]. In this Opinion article, we make use of the Alexander–Briggs notation to characterize knots (e.g., as in Figure 2A). In this notation, a knot is represented by two numbers:

[1]Medical Systems Biophysics and Bioengineering, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands
[2]Department of Synthetic Biology and Immunology, National Institute of Chemistry, Ljubljana, Slovenia
[3]Department of Macromolecular Biochemistry, Leiden Institute of Chemistry, Leiden, The Netherlands

*Correspondence:
a.mashaghi.tabari@lacdr.leidenuniv.nl
(A. Mashaghi).

https://doi.org/10.1016/j.trechm.2020.04.009

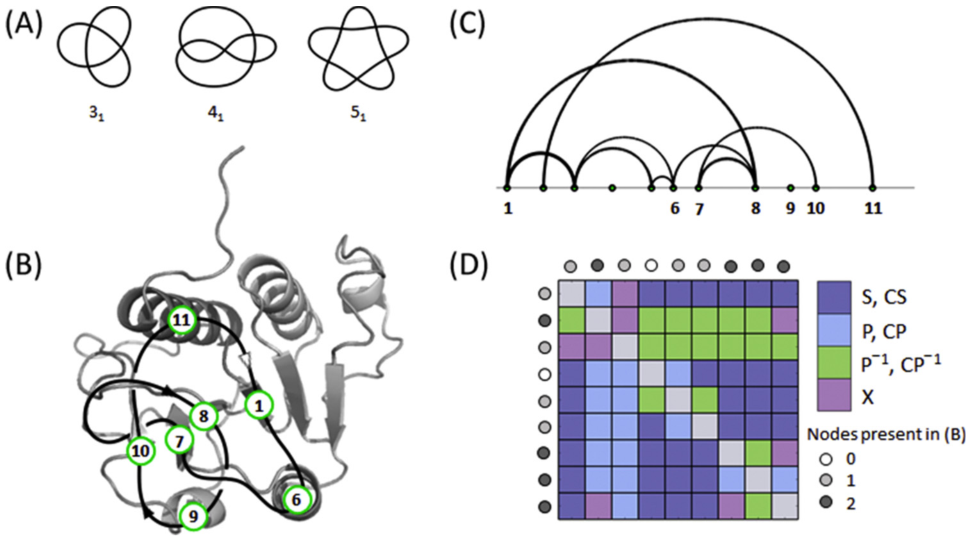


Trends in Chemistry

Figure 1. Simplicial Complex Representation of the Circuit Topology Formalism. (A) Simplicial complex representation of two contacts in parallel (P), series (S), and cross (X) relation. To transition from one configuration to another, vertices and nodes need to be edited (added or removed). Such analysis provides a framework for the calculation of distances between structures via a graph editing approach. In (B) a further example of this concept is shown, applied to the folding process: the folding of the E adenine riboswitch is represented in a graph where the x-axis represents the graph edit distance and the y-axis the number of contacts (see [13] for further information).

the main one indicates the crossing number and its subscript provides the identification number of the knots with the same crossing number.

A folded chain is formed when a polymer establishes intrachain contacts: two contact sites along the chain come in close proximity, creating either pairwise or higher-order connections. The CT of a folded polymer chain defines the arrangement of intramolecular contacts with respect to the



Trends in Chemistry

Figure 2. Knot and Circuit Topology Representation: A Comparison. Topology representations of the YibK methyltransferase, which exhibits a 3_1 knot forming the cofactor binding site. (A) Projections of three knots. YibK is an example of a 3_1 or trefoil knot. (B) Protein structure, with knot diagram overlaid. Secondary structural elements along the knot diagram are numbered according to their position along the backbone. (C) Circuit topology diagram with numbered elements in (B) numbered underneath the diagram. Cutoff: 3.7 Å, four contacts. (D) Circuit topology matrix of YibK methyltransferase, retrieved from the diagram in (C). The grayscale dots represent the numbers of nodes for each interacting loop pair that are part of the knot.

path between polymer ends. The approach is simple and generic and can be represented according to an algebraic formalism, providing quantitative measures for comparative analysis and experimental studies. For a given pair of binary contacts, three arrangements can be identified: parallel (P), series (S), and cross (X) (Figure 1A) [5]. In addition, two contacts may be in concerted series (CS) or parallel (CP) relation if they share a site [7]. Here, a 'site' or node may have one of several definitions: for instance, it may be a protein residue, a single nucleotide, or an element of secondary structure. Furthermore, the definition of a contact may incorporate a cutoff distance (or atom-type-specific distances) and number of atom–atom contacts or may focus on a particular type of contact such as a disulfide bond. When needed, one can simplify the representation; for example, by treating the asymmetric parallel relation as a symmetric one or by extending the definitions of P and S to merge them with CP and CS relations, respectively. Given suitable definitions, a matrix of relations between pairs of contacts may be constructed.

Here, we focus on CT and geometric topology approaches or, more specifically, on knot theory. There is a wide variety of other topological methods that have been developed for molecular sciences, including algebraic topology (e.g., persistent homology) and differential topology (e.g., de Rham Hodge theory, quantum topology, topological order) [8]. Among these methods, persistent homology appears to be more promising for biomolecules [9–12]. Persistent homology approaches are reviewed elsewhere [8]. However, we note that CT analysis can be readily combined with existing persistent homology tools. The CT motifs introduced earlier can be readily represented in the form of simplicial complexes and subjected to algebraic topology analysis (Figure 1) [13].

Topological Analysis of Proteins

Proteins, known as the primary machinery of life [14], often need to fold transiently or permanently into one or more specific spatial conformations, mostly driven by noncovalent interactions

[15,16]. Among the unlimited possibilities of arrangements, a limited number of motifs and domains is exhibited by nature, evidencing some general rules that govern the complexities of protein structure [17]. Various theoretical methods, including knot theory [18] , knotoids [19], and, recently, CT [5], have been developed to formalize the structural relationships among diverse proteins.

Knots and Knotoids in Proteins

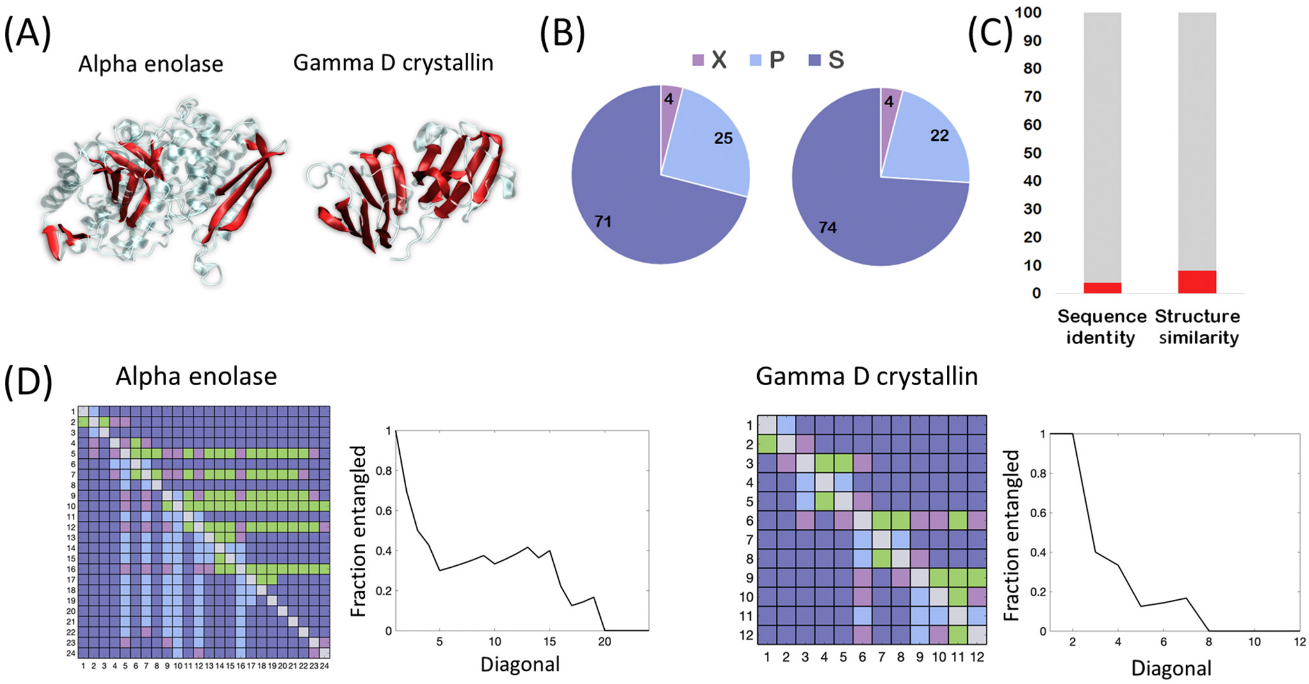
Structural analysis of 400 knotted proteins, diverse in sequence and family, showed that the knotting pattern in proteins is strictly evolutionarily conserved [20]. There are several families of proteins that reproducibly form simple knots, complex knots, and slipknots; in these proteins, the disadvantage of less efficient folding may be balanced by a functional advantage connected with the presence of these knots [21]. Knot theory appears to be a powerful approach to explore the structural, mechanical, and functional roles of such entangled topological features in proteins [22]. Most of the knots are located in functionally important positions in protein structure. Recent research has established that knotted cores, and especially their borders, show strong enrichment in the number of contacts with surrounding structural elements. Buried inside the protein structure, these regions showed increased thermal stability, providing a favorable environment for the protein active site [23]. Despite advances, knot theory has limitations. The fraction of knotted proteins is only 0.77% of all proteins [24]. Also, to adhere with formal mathematical definitions, knots must be closed rings, which are rare in protein structures. In 2012, as a generalization of knot theory, knotoids were introduced as diagrams representing projections of open curves in 3D space [25]. Due to the open and dynamic nature of protein structure, knotoids have attracted interest for studies of global and local entanglements of proteins [26]. Results are now accessible through online databases [27,28].

CT of Proteins

Despite many applications, both knot and knotoids theories ignore intrachain interactions. CT is well suited to address this challenge. Within this framework, a wide range of biologically important interactions could be considered: for example, contacts inclusive of the covalent S–S bond, interactions between secondary structural elements (e.g., β–β, α–α), and connections between coevolving groups known as sectors [29]. These connections could be extracted from solved 3D structures or even determined from state-of-the-art single-molecule force spectroscopy measurements. Force measurements on model proteins and human steroid receptor proteins demonstrated the different steps of conformational change at distinct lengths [30,31]. Force jumps between different lengths could be related to the breaking of connections. CT aims to provide a framework to gain molecular insight (including allowed and forbidden folding transitions) from force spectroscopy data; that is, force-versus-length diagrams [32].

The native CT of a protein may inform on its function. A fundamental question in biology is whether the function of a protein is correlated with topology. For instance, crystallin is a moonlighting protein [33] mainly known as a structural protein but also, in some cases, exhibiting enolase activity. Comparison of the CT maps of α-crystallin and human enolase showed negligible sequence and geometric similarities (Figure 3A–C) but striking similarities in the frequencies of topology motifs extracted from the CT matrix (Figure 3D). Future studies will show whether topological similarities are generically associated with functional similarities.

CT provides insight into folding mechanisms. It has been shown that among various structural descriptors of proteins, contact order, size, and CT are folding rate predictors [5]. Being size invariant and flexible in defining the contacts, CT has advantages over contact order in estimating the folding rates and number of unfolding paths of a macromolecule [3]. When other determinants

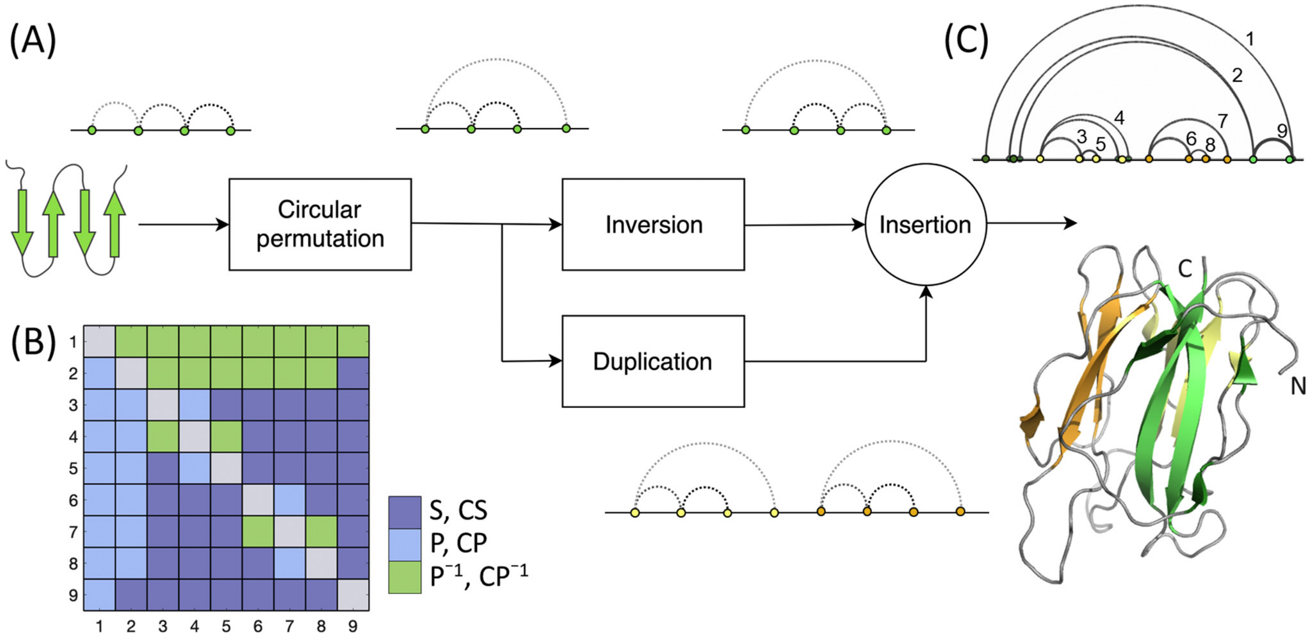


Trends in Chemistry

Figure 3. Topological Comparison between Two Moonlighting Proteins. Preliminary example data showing two proteins that are different in sequence and structure, yet similar in topology and function. (A) Crystal structure of α-enolase and gamma crystallin D: red parts indicate extended beta strands. (B) Striking similarity in the frequencies of topology motifs extracted from a circuit topology matrix for atom–atom topologies. P, parallel; S, series; X, crossing arrangements. (C) The two proteins have negligible sequence overlap and structural similarity as estimated by the jFATCAT_rigid comparison method. (D) Circuit topology matrices for the two proteins: the 'entangled' relations (i.e., parallel and cross) are clustered along the diagonal and decay with distance from the diagonal in a similar way for each protein. Our analysis raises the question of whether these topological similarities can be generically related to protein function.

of folding rates (e.g., size, contact order) are similar for two molecules, CT can readily resolve differences in folding kinetics [3]. Folding pathways can be mapped onto a topology landscape, allowing the identification of topological transitions and topological traps (misfolds) [34]. The topology concept has also been used to define simple reaction coordinates to illustrate the progress of conformation-dependent reactions [13,34]. Furthermore, CT analysis reveals how transient interactions with molecular chaperones guide the folding process towards certain topologies and away from others [35,36]. The recent development of single-molecule techniques to study protein folding has led to an increased demand for theoretical tools to interpret experimental data [37,38]. CT analysis combined with other molecular modeling approaches may reveal working principles of molecular chaperones [32,39].

The CT of proteins may change on rearrangement of the corresponding genetic materials. Recently, Schullian and colleagues developed a mathematical framework to describe the CT of a biomolecule and topological changes such as standard and circular permutation, duplication, and the addition/elimination of contacts [7]. It was found that topology permutations underlie aspects of protein evolution and dynamics such as domain swapping on mutation and hairpin flipping within a beta barrel. Figure 4 shows how a relatively complex protein can be built from permutations and combinations of a simple topology. Figures 4A and 3C show the progression from a simple topology to the final product, for the protein membrane protein VMO-I. Figure 4C shows the topology matrix of the protein. Related experimental studies have revealed the implications of such molecular engineering operations for folding dynamics. For example,



Trends in Chemistry

Figure 4. Molecular Operations as Topological Permutations. A more complex protein topology is built from permutations of a simple circuit topology motif. (A) Diagram showing the construction of the circuit topology of membrane protein VMO-I (PDB ID 1VMO) based on permutations of the concerted series arrangement belonging to the up-down-up-down four-strand motif. (B) Relations between contacts in the circuit topology of VMO-I, with sites numbered as in (C). P^{-1} indicates the inverse of the parallel relation (loop *i* includes loop *j*), while CP and CS are parallel and series relations in which one of the contact sites is shared between the two loops (see [5]). (C) Circuit topology structure of VMO-I, with nodes corresponding to beta-strand segments and edges weighted according to the number of contacts. An atom–atom distance cutoff of 3.5 Å and a number-of-contacts threshold of five contacts were used.

single-molecule experiments with optical tweezers showed that circular permutation of the amino acid sequence of the T4 lysozyme greatly affects its folding cooperativity, indicating that protein topology could play a role in preventing partial unfolding and subsequent misfolding [40]. The full relevance of the CT framework to molecular evolution and dynamics remains an active area of study.

Knot theory applications also took a leap forward with the efforts by Adams and colleagues [41], extending classical knot theory to intrachain contacts. CT can also be extended to describe both contacts and knot crossings (unpublished data). These extensions show promise to promote a unification of knot-based topology framework and contact-based CT framework, providing a uniform language for a working topological description. These two approaches can provide complementary descriptions of a polymer, such as a protein (Figure 2). Figure 2B shows the knotted protein YibK, with the overlaid knot diagram of the knotted portion of the structure alongside the protein topology diagram and matrix (Figure 2C,D). Strands 1, 8, 7, and 10 form part of the protein's beta sheet, including two contacts in cross relation.

Topological Analysis of Nucleic Acids

Cellular nucleic acids often fold into globular structures to achieve function. Folding happens at various scales, from small RNA molecules to large eukaryotic genomes. Various topological concepts, including supercoiling, knot theory, and contact arrangement, have been developed to describe folded nucleic acids. In what follows, we summarize these developments and discuss how CT can be used as a universal topology framework.

Topology of RNA

RNA molecules may fold back on themselves to form complex 3D shapes capable of ligand/target recognition and catalysis. These structures can be achieved by means of several mechanisms, including: hydrogen bonding and stacking interactions as in tRNA tertiary structure; ions bound to specific sites, as found in rRNA fragments; and pseudoknot folds, as seen in mRNA fragments with extensive noncanonical pairing structure (in contrast to canonical Watson–Crick pairing) [42]. Pseudoknots, which are segments in the secondary structure where half of one stem is intercalated between the two halves of another stem, are abundant in RNA molecules and can have important functional implications [43]; thus, topological classifications of RNA have been mainly focused on pseudoknots and on the concept of topological genus. RNA secondary structure can be schematically represented by a planar diagram, with straight lines representing the backbone of the molecule and arches representing the bonds that give the molecule its characteristic folded shape. The RNA secondary structure is said to be pseudoknotted if the diagram indicates crossing among base pairs (Figure 5C,D). These crossings in the diagram are equivalent to cross relations in the CT framework [44]. Figure 5A,B displays RNA structures that are not pseudoknotted (no cross relations). A given RNA molecule that has been thus represented can then be characterized by the genus of the auxiliary 2D surface associated with the diagram; that is, a sphere with handles. The genus *g* of a diagram is the minimum number of handles a sphere must have to enable drawing of the diagram on it without any crossing [44–47].

By comparing the planar diagram representation with the CT diagram (Figure 5), it is possible to draw a parallel between planar diagrams and CT: cross relations, where the number of arches *n* is equal to the number of loops in the chain. Therefore, if we were to calculate the topology matrix of

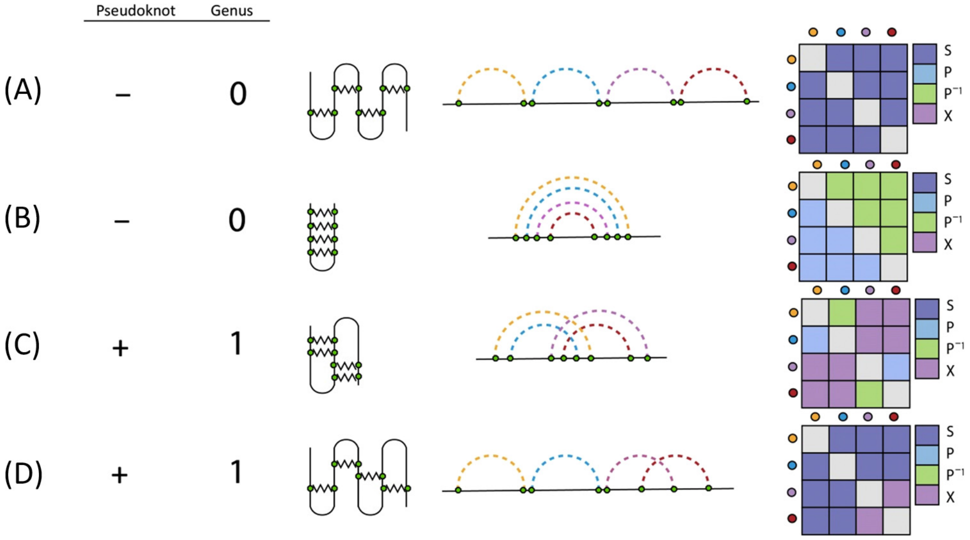


Figure 5. Pseudoknots, Genus, and Circuit Topology (CT): A Comparison. Four examples of RNA structures and their diagrammatic representations. The first two structures (A,B) are not pseudoknotted, while structures (C,D) contain pseudoknots. Pseudoknots correspond to cross relations in the CT matrix. We show that structures with the same genus can have dramatically different topologies. The genus is a positive integer that quantifies the topological complexity of the diagram and therefore of the folded RNA structure [44]. Structure (A) has genus 0 but contains only series relations. Structure (B), by contrast, while still having genus 0, contains only the so-called ‘entangled’ relations: parallel and cross. Similarly, (C,D) have genus 1, although structure (C) contains only cross and parallel relations and structure (D) is dominated by series relations. The only common trait in (C,D) is the presence of cross relations, which indicate the pseudoknot.

a pseudoknotted RNA molecule, we would obtain an $n \times n$ matrix, such as the one represented in Figure 5. Given the similarity between these two representations, the genus of a CT diagram can be readily calculated, and consequently a topology matrix. Recently, pseudoknot classification and comparison in RNA molecules was given a new algebraic formalism [48]. Here, RNA structures are represented as expressions of an algebraic language with three operators (concatenation, nesting, and crossing) and simple hairpin loops as operands. In the language of CT, concatenation, nesting, and crossing correspond to series, parallel and cross relations. These relations were also given an operator representation [7]. Other creative frameworks exploit graph theory. The RNA-as-graph approach involves the translation of an RNA 2D structure into tree and dual graph objects. In tree graphs, stems are the edges, while junctions, bulges, and loops are the vertices [49]. Once again, we draw a parallel with CT, in its simplicial complex representation [13], where transitions from one topology relation to another are described and quantified in terms of graph editing (e.g., addition or removal of vertices and edges from a graph) (Figure 1B). CT thus provides a unified language for the description of RNA folds.

Topology of Cellular DNA

Genome topology plays a fundamental role in the regulation of gene expression. The genomic spatial arrangement is shaped by chromatin long-range interactions, which are mediated by architectural proteins such as CTCF and cohesin [50,51]. These interactions cause the chromatin to form loops and eventually organize in topologically associating domains (TADs) in mammals. Whereas the molecular mechanistic basis for loop formation may differ, similar types of domain arrangements are found in lower eukaryotes [50] and prokaryotes [52,53]. Alterations in chromatin topology are key to cell differentiation [54] and have been implicated as drivers of oncogenic programs [55]. Genetic mutations that affect chromatin topology potentially lead to changes in gene expression, therewith facilitating disease susceptibility and evolutionary adaptation [56]. Providing a rigorous topological framework is therefore a fundamental step to shed light on the link between genome topology and function.

Supercoiled DNA

Early efforts to characterize topology in DNA were focused on supercoiling. DNA supercoiling is the consequence of twisting DNA: it describes the coiling of the axis of the double helix. Supercoiling occurs in the DNA of organisms at all levels of evolutionary complexity. Human interphase chromosomes are divided into domains with different levels of supercoiling, where under-wound domains are transcriptionally active, cytologically decondensed, and topologically constrained [57]. These domains were shown to frequently correspond to TADs detected by chromosome conformation capture (3C) methods [57]. A topological constant commonly used to characterize supercoiling is the linking number Lk, which represents the number of times the two strands of the DNA double helix are intertwined [58]. This parameter can be expressed as the sum of two geometric parameters: writhe (Wr) and twist (Tw) [58]. Wr measures the coiling of the DNA axis and Tw the helical winding of the DNA strands around each other. Although this formalism was developed for circular DNA, supercoiling has also been observed and studied experimentally in linear segments of double-stranded DNA (dsDNA) [59,60]. The twisted linear DNA forms intertwined loops called plectonemes, the dynamics of which could be studied as they diffuse or hop along the DNA strand. From formal point of view, plectonemic loops and their dynamics can be readily represented with CT terminology.

Knotted DNA

DNA at short length scales (<50 nm) is a stiff polymer, but its considerable length – of the order of millimeters in bacteria and meters in humans – makes it very liable to self-entanglement and knotting [61]. Knots in packaged viral DNA have been widely documented in the literature [62]. The micron-long viral DNA molecules are tightly packed and condensed inside a capsid, which is about 50–80 nm in size [35]. This strong confinement facilitates the occurrence of knots, with a distribution of knot types that is biased towards complex knots: gel electrophoresis characterization revealed a predominance of the torus knot 5_1 and scarcity of the achiral knot 4_1 [63]. More recently, small steady-state fractions of DNA knots were also found in chromatin inside cells [64]. There is debate about the extent and scale to which knots are present at the chromosome scale. The 100-kb resolution analysis of individual chromosomes in the nuclei of single haploid mouse embryonic stem (ES) cells obtained by Hi-C contact data [65] revealed that chromosomes do contain knots, with the fraction of unknotted chromosomes being less than 20% [66]. Moreover, knots with more than five crossings or even multiple knots appeared to be the most common kind, representing more than 50% of the knot population [66].

Various single-molecule techniques have also been used to characterize DNA knots. For instance, nanopore sensors have been used to map the equilibrium configurations of DNA knots, revealing a wide distribution in tightness. The persistence of very loose knots might have implications for understanding the efficiency of the biological mechanisms accountable for unknotting the molecules [67], like, for example, the action of type II DNA topoisomerases [68]. Considering the new wealth of information we have on contacts in genomic structures (see later) and the high likelihood of these structures producing complex knots, a generalized knot theory approach such as the one presented by Adams and colleagues [41] may be a useful direction for future research.

Contact Arrangement in Cellular DNA

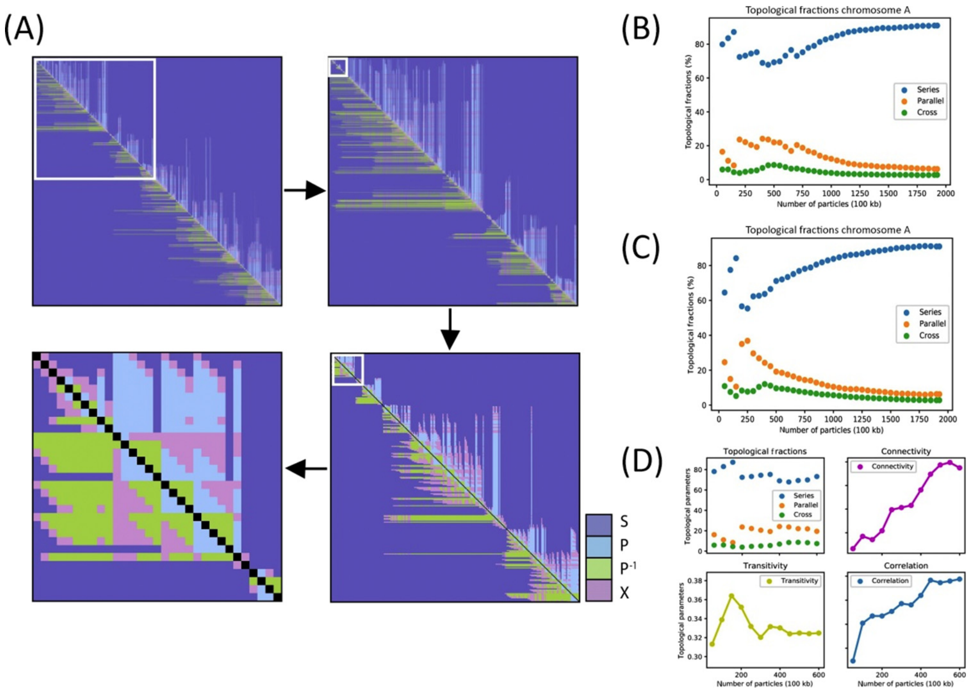
The development of innovative technologies such as fluorescence *in situ* hybridization, *in vivo* tagging of genomic loci, and 3C- and Hi-C-based technologies have led to an increase in the available structural information. Hi-C technology is particularly suited as a source of topological data for chromosomes since it allows the identification of long-range interactions in a genome-wide fashion [69]. This process results in large libraries of pairwise chromatin interactions, which reveal highly reproducible features such as TADs [70]. CT could serve as novel framework to analyze these data, since it provides a contact-based description of topology. Hi-C contact maps can easily be used to derive topology matrices, which can add complementary topological information (Figure 6). Figure 6A shows progressive close-ups into the CT matrix of a chromosome; in a conceptually similar fashion, Figure 6B,C shows the topological fractions for chromosome sections of different sizes, while presenting a comparison with parameters derived from network topology (Figure 6D). Moreover, CT has already been used to describe polymers that fold under confinement [35], which is the case for the DNA encapsulated in the cell or inside its organelles. Also, a formalism such as the one described in [35] could be used to describe the temporal evolution of genomic domains highlighted by single cell Hi-C data.

Topology of Organic and Bioinspired Polymers

Advances in molecular-engineering-enabled synthesis of molecular knots and topological polymers have led the way towards applications in several fields, including chemical biology, medicine, and materials science.

Engineered Folded DNA Structures

DNA has been demonstrated as a versatile building block for objects such as 2D crystals [71], nanotubes [72], and 3D nanopolyhedra [73]. Many DNA-based materials involve branched molecules (DNA bricks), in which branch points represent the vertices of various types of polyhedra [73,74]. These building blocks are created with techniques that combine hybridization



Trends in Chemistry

Figure 6. Circuit Topology (CT) Analysis of a Chromosome. CT analysis of the first chromosome of a single mouse embryonic stem (ES) cell [65]. (A) CT matrix calculated by choosing a cutoff radius (two particle radii) to define contacts between 100-kb particles. The four images show progressive close-ups on smaller areas of the matrix. We can see how parallel and cross relations cluster around the diagonal, suggesting a domain-like structure. In (B,C) we can see the calculated topological fractions for progressively higher numbers of particles. This cumulative analysis shows how we can have an indication of the domain-like structures when we have a high resolution (few particles). When the resolution lowers (right side of the graph), the fractions remain constant, indicating that the relative proportions of series, cross, and parallel do not depend on the number of kilobases included in the analysis after a certain threshold (which for this chromosome is about 750 particles). The particles were computed in (B) from the top of the PDB file to the bottom and in (C) from the bottom up. (D) Comparison between CT analysis and network analysis for the first 600 particles. The network was built using contact sites as nodes and contacts as edges. Two network parameters – average connectivity and Pearson correlation – show stepwise behavior similar to that displayed by the topological fractions. Transitivity exhibits a rise at small scales.

(sticky-ended cohesion) and synthetic stable branched DNA (e.g., as Holliday junctions). Others rely on the design of scaffolded DNA origami, where one long, single-stranded DNA molecule is folded into arbitrary 2D shapes, which are then the building blocks for larger assemblies [75]. Here, we focus on those cases where a single molecule is folded.

There is a growing interest in designing knotted nucleic acids [76]. Kočar and colleagues presented the design principles to fold highly knotted single-chain DNA nanostructures. One of the key principles hereby demonstrated is the identification of favorable and unfavorable folding steps, from a topological and kinetic point of view. These steps are identified by considering the pairwise connections that are created during folding and by classifying these connections using CT [5]. This strategy demonstrated that highly knotted structures can be formed based on the stepwise formation of connections defined by their decreasing stability as the alternative folding pathways that results in structures of the same stability could not form the knotted structures. This is an example of how CT and knot theory combined can be used to engineer the topological features of a chain.

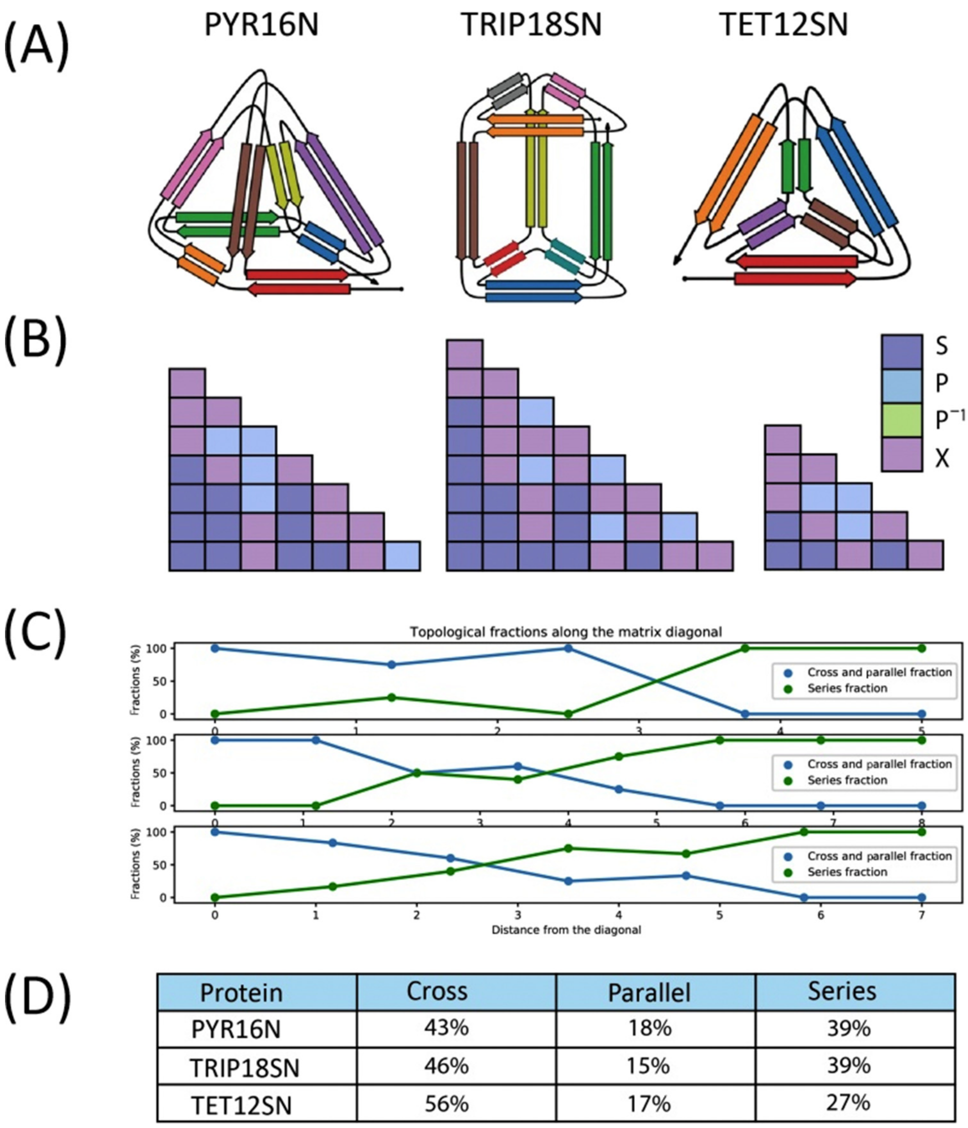
Concerning contact-based topology, Han and colleagues presented a new strategy to design and synthesize a single DNA or RNA strand to self-fold into a complex (user-prescribed) structure [77]. In their approach, single molecules of ssDNA and RNA with synthetic sequences ranging in length from ~1000 to ~10 000 nt were folded into origami. Knotting in these structures is prevented to avoid kinetic traps and assure smooth folding. While these origami structures are unknotted, their contact arrangement topology is quite elaborate. The design principle is based on parallel crossover, with layers that are covalently linked in a raster filling pattern. From a knot perspective, these structure all belong to the same class (the unknot). Therefore, a contact base topology such as CT is necessary to detect their distinguishing features.

Synthetic Proteins

Advances in nucleic acid engineering have inspired analogous designs for proteins. Proteins are programmable polymers, which can fold into elaborate 3D structures and are therefore particularly versatile for the engineering of materials with tailormade structure and function. Design principles correlating loop geometries and secondary structure packing orientation allow accurate protein size and length control, as investigated by Baker and colleagues [78,79,81]; this loop-based characterization is highly compatible with CT. Ljubetic and colleagues designed self-assembling coiled-coil protein-origami (CCPO) cages of various geometries (tetrahedra, a four-sided pyramid, and a triangular prism) and provided a computational platform for the design of arbitrary complex CCPO polyhedra [80]. These structures combine the modular building strategy of DNA (DNA bricks) nanotechnology with the programmable functionality of amino acids. They have interesting physical properties, which can be studied from a circuit topological point of view. In Figure 7A, we show the CCPO cage structures from [80] and the corresponding CT matrices (Figure 7B). We can see in Figure 7C,D that these three cages are strikingly similar with regard to topological fractions (percentages of series, cross and parallel relations) and show relatively low contact order. This proof-of-concept study demonstrates that the CCPO cages can be constructed with desired contact order and topology. We note that topology determines folding pathway; furthermore, the topology and contact order may independently affect the folding rate. However, a systematic analysis of CCPO cages based on topological traces with different CT but shared contact order has not been performed, and whether some combinations of the circuit topological fractions might be more helpful than others in promoting stability and other kinetic properties remains to be seen. While knots could potentially form between linker regions located at the vertexes of CCPO polyhedra, they cannot be programmed into the designs at the current stage. Extending the length of CC building modules to encompass a full turn has potential for the design of knotted protein structures with the possibility of designing the folding pathway and making highly knotted proteins or polypeptide-based materials.

Topology and Organic Chemistry

New topological features at the molecular level can introduce new material properties. Many efforts in this direction have been focused on molecules created by interlocked chains (as opposed to single folded chains), such as catenanes and rotaxanes [82], and on networks of interconnected molecules called polymer networks [83]. However, the field is also starting to obtain a better understanding of the strategies needed for the synthesis of a single entangled molecular strand, as in the case of molecular knots. The steric restrictions imparted to the molecule by knotting hinder the range of movement of the molecular components, significantly influencing physicochemical properties [84]. So far, four types of knots have been successfully synthesized using small-molecule building blocks: the trefoil [85,86], the figure-eight [87], the pentafoil [88], and the 8_{19} knot (a knot with eight crossings) [89]. A comprehensive theoretical


(A)
PYR16N
TRIP18SN
TET12SN
(B)
S
P
P^{-1}
X
(C)
Topological fractions along the matrix diagonal
Fractions (%)
100
50
0
Cross and parallel fraction
Series fraction
0
1
2
3
4
5
6
7
8
Distance from the diagonal
(D)

Protein	Cross	Parallel	Series
PYR16N	43%	18%	39%
TRIP18SN	46%	15%	39%
TET12SN	56%	17%	27%

Trends in Chemistry

Figure 7. Circuit Topology (CT) Analysis of Origami Proteins. (A) Three examples of coiled-coil protein-origami (CCPO) cages [80]: namely, a tetrahedron (TET12SN), a pyramid (PYR16N), and a trigonal prism (TRIP18SN). (B) The CT matrices (of which half are shown, since they are symmetrical) show remarkable similarities both in the percentages of topological fractions (D) and in how they are located in the matrix. First, most of the dominant topological fractions appear to be cross in all three cases. Second, most of the parallel and cross relations are clustered along the diagonal of the matrix, indicating that most short-range contacts have this type of arrangement. Series contacts are present only in the corner of the matrix, indicating that series dominates long-range distances along the chain. If we calculate the percentages of the fractions along diagonal lines in the matrix from the matrix diagonal ($i = j$) towards the periphery of the matrix and plot the percentages (C), we see that cross and parallel start from a maximum and decrease to zero while series has a different behavior, starting from zero and reaching a maximum at maximum distance from the diagonal.

framework would not only allow characterization but would also be beneficial for the practical purpose of purifying polymers with different topologies, as exemplified in [90]. In that study, it was shown using simulations how nanopores can be used to sense and enrich certain circuit topologies.

Concluding Remarks and Future Perspectives

Contact-based CT and knot theory form two complementary frameworks for describing, understanding, and engineering linear biopolymers such as proteins and nucleic acids, as summarized in the Outstanding Questions. An important future development will be further integration of these two applied theories and the establishment of how they can be more generally utilized in prediction and design. Towards this goal, it is likely that machine learning and artificial intelligence (AI), including recent advances in neural networks, will play key roles, perhaps paralleling major recent successes in the application of persistent homology and machine learning to biomolecular analysis and discovery [91].

Outstanding Questions

Contact-based circuit topology and knot theory form two complementary frameworks for describing and understanding the topology of folded biopolymers. Can a unified theory be developed that includes both contacts and crossings (knots)? This would be a great asset towards a comprehensive description and applications.

Will the availability of new single-cell Hi-C genomic data fuel contact-based topological analysis of these datasets? Circuit topology could represent an ideal framework for this analysis and help to characterize interchromosomal and intercellular heterogeneity.

Can we develop a comprehensive theoretical framework that includes molecular loops and entanglement to uncover the full potential of topology engineering for the creation of new material properties?

What role will machine learning and AI, including recent advances in neural networks, play in topological prediction and design?

References

1. Wang, X.-W. and Zhang, A.-B. (2018) Chemical topology and complexity of protein architectures. *Trends Biochem. Sci.* 43, 806–817
2. Dabrowski-Tumanski, P. and Sulkowska, J.I. (2017) Topological knots and links in proteins. *Proc. Natl. Acad. Sci. U. S. A.* 114, 3415–3420
3. Mugler, A. *et al.* (2014) Circuit topology of self-interacting chains: implications for folding and unfolding dynamics. *Phys. Chem. Chem. Phys.* 16, 22537–22544
4. Flapan, E. *et al.* (2019) Topological descriptions of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 116, 9360–9369
5. Mashaghi, A. *et al.* (2014) Circuit topology of proteins and nucleic acids. *Structure* 22, 1227–1237
6. Faísca, P.F.N. (2015) Knotted proteins: a tangled tale of structural biology. *Comput. Struct. Biotechnol. J.* 13, 459–468
7. Schullian, O. *et al.* (2020) A circuit topology approach to categorizing changes in biomolecular structure. *Front. Phys.* 8, 5
8. Nguyen, D.D. *et al.* (2020) A review of mathematical representations of biomolecular data. *Phys. Chem. Chem. Phys.* 22, 4343–4367
9. Cámara, P.G. (2017) Topological methods for genomics: present and future directions. *Curr. Opin. Syst. Biol.* 1, 95–101
10. Xia, K. and Wei, G.-W. (2014) Persistent homology analysis of protein structure, flexibility, and folding. *Int. J. Numer. Method. Biomed. Eng.* 30, 814–844
11. Cang, Z. and Wei, G.-W. (2017) TopologyNet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* 13, e1005690
12. Nguyen, D.D. *et al.* (2019) Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *J. Comput. Aided Mol. Des.* 33, 71–82
13. Verovšek, S.K. and Mashaghi, A. (2016) Extended topological persistence and contact arrangements in folded linear molecules. *Front. Appl. Math. Stat.* 2, 6
14. Brocchieri, L. (2005) Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* 33, 3390–3400
15. Mollica, L. *et al.* (2016) Binding mechanisms of intrinsically disordered proteins: theory, simulation, and experiment. *Front. Mol. Biosci.* 3, 52
16. Kim, D.-H. and Han, K.-H. (2018) Transient secondary structures as general target-binding motifs in intrinsically disordered proteins. *Int. J. Mol. Sci.* 19, 3614
17. Tubiana, J. *et al.* (2019) Learning protein constitutive motifs from sequence data. *Elife* 8, e39397
18. Mishra, R. and Bhushan, S. (2012) Knot theory in understanding proteins. *J. Math. Biol.* 65, 1187–1213
19. Goundaroulis, D. *et al.* (2017) Topological models for open-knotted protein chains using the concepts of knotoids and bonded knotoids. *Polymers (Basel)* 9, E444
20. Sułkowska, J.I. *et al.* (2012) Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Natl Acad. Sci. U. S. A.* 109, 1205918109
21. Sulkowska, J.I. *et al.* (2012) Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Natl. Acad. Sci. U. S. A.* 109, E1715–E1723
22. Patil, V.P. *et al.* (2020) Topological mechanics of knots and tangles. *Science* 367, 71–75
23. Dabrowski-Tumanski, P. *et al.* (2016) In search of functional advantages of knots in proteins. *PLoS One* 11, e0165986
24. Alexander, K. *et al.* (2017) Proteins analysed as virtual knots. *Sci. Rep.* 7, 42300
25. Vladimir, T. (2012) Knotoids. *Osaka J. Math.* 49, 195–223
26. Goundaroulis, D. *et al.* (2017) Studies of global and local entanglements of individual protein chains using the concept of knotoids. *Sci. Rep.* 7, 6309
27. Dabrowski-Tumanski, P. *et al.* (2019) KnotProt 2.0: a database of proteins with knots and other entangled structures. *Nucleic Acids Res.* 47, D367–D375
28. Dorier, J. *et al.* (2018) Knoto-ID: a tool to study the entanglement of open protein chains using the concept of knotoids. *Bioinformatics* 34, 3402–3404
29. Halabi, N. *et al.* (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell* 138, 774–786
30. Suren, T. *et al.* (2018) Single-molecule force spectroscopy reveals folding steps associated with hormone binding and activation of the glucocorticoid receptor. *Proc. Natl. Acad. Sci. U. S. A.* 115, 11688–11693
31. Mashaghi, A. *et al.* (2014) Misfolding of luciferase at the single-molecule level. *Angew. Chem. Int. Ed.* 53, 10390–10393
32. Heidari, M. *et al.* (2019) Mapping a single-molecule folding process onto a topological space. *Phys. Chem. Chem. Phys.* 21, 20338–20345
33. Huberts, D.H.E.W. and van der Klei, I.J. (2010) Moonlighting proteins: an intriguing mode of multitasking. *Biochim. Biophys. Acta* 1803, 520–525
34. Mashaghi, A. and Ramezanpour, A. (2015) Distance measures and evolution of polymer chains in their topological space. *Soft Matter* 11, 6576–6585
35. Satarifard, V. *et al.* (2017) Topology of polymer chains under nanoscale confinement. *Nanoscale* 9, 12170–12177
36. Heidari, M. *et al.* (2017) Topology of internally constrained polymer chains. *Phys. Chem. Chem. Phys.* 19, 18389–18393
37. Mashaghi, A. *et al.* (2013) Reshaping of the conformational search of a protein by the chaperone trigger factor. *Nature* 500, 98–101
38. Mashaghi, A. *et al.* (2016) Alternative modes of client binding enable functional plasticity of Hsp70. *Nature* 539, 448–451
39. Singhal, K. *et al.* (2015) The trigger factor chaperone encapsulates and stabilizes partial folds of substrate proteins. *PLoS Comput. Biol.* 11, e1004444
40. Shank, E.A. *et al.* (2010) The folding cooperativity of a protein is controlled by its chain topology. *Nature* 465, 637–640
41. Adams, C. *et al.* (2019) Knot theory for proteins: gauss codes, quandles and bondles. *arXiv.* Published online December 19, 2019. http://arxiv.org/abs/1912.09353
42. Draper, D.E. (1996) Strategies for RNA folding. *Trends Biochem. Sci.* 21, 145–149
43. Giedroc, D.P. and Cornish, P.V. (2009) Frameshifting RNA pseudoknots: structure and mechanism. *Virus Res.* 139, 193–208
44. Bon, M. *et al.* (2008) Topological classification of RNA structures. *J. Mol. Biol.* 379, 900–911
45. Ringeisen, R.D. (1979) Survey of results on the maximum genus of a graph. *J. Graph Theory* 3, 1–13
46. Zając, S. *et al.* (2018) Genus trace reveals the topological complexity and domain structure of biomolecules. *Sci. Rep.* 8, 17537
47. Rubach, P. *et al.* (2020) Genus for biomolecules. *Nucleic Acids Res.* 48, D1129–D1135

48. Quadrini, M. *et al.* (2019) An algebraic language for RNA pseudoknots comparison. *BMC Bioinformatics* 20, 161
49. Schlick, T. (2018) Adventures with RNA graphs. *Methods* 143, 16–33
50. Cubeñas-Potts, C. and Corces, V.G. (2015) Architectural proteins, transcription, and the three-dimensional organization of the genome. *FEBS Lett.* 589, 2923–2930
51. Dekker, J. and Misteli, T. (2015) Long-range chromatin interactions. *Cold Spring Harb. Perspect. Biol.* 7, a019356
52. Dame, R.T. and Tark-Dame, M. (2016) Bacterial chromatin: converging views at different scales. *Curr. Opin. Cell Biol.* 40, 60–65
53. Dame, R.T. *et al.* (2019) Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nat. Rev. Genet.* 21, 227–242
54. Stadhouders, R. *et al.* (2019) Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* 569, 345–354
55. Flavahan, W.A. *et al.* (2019) Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. *Nature* 575, 229–233
56. Tang, Z. *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 163, 1611–1627
57. Naughton, C. *et al.* (2013) Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat. Struct. Mol. Biol.* 20, 387–395
58. White, J.H. *et al.* (1988) Helical repeat and linking number of surface-wrapped DNA. *Science* 241, 323–327
59. van Loenhout, M.T.J. *et al.* (2012) Dynamics of DNA supercoils. *Science* 338, 94–97
60. Vlijm, R. *et al.* (2015) Experimental phase diagram of negatively supercoiled DNA measured by magnetic tweezers and fluorescence. *Nanoscale* 7, 3205–3216
61. Shaw, S. and Wang, J. (1993) Knotting of a DNA chain during ring closure. *Science* 260, 533–536
62. Liu, L.F. *et al.* (1981) Knotted DNA from bacteriophage capsids. *Proc. Natl. Acad. Sci. U. S. A.* 78, 5498–5502
63. Arsuaga, J. *et al.* (2005) DNA knots reveal a chiral organization of DNA in phage capsids. *Proc. Natl. Acad. Sci. U. S. A.* 102, 9165–9169
64. Valdés, A. *et al.* (2018) DNA knots occur in intracellular chromatin. *Nucleic Acids Res.* 46, 650–660
65. Stevens, T.J. *et al.* (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544, 59–64
66. Siebert, J. *et al.* (2017) Are there knots in chromosomes? *Polymers (Basel)* 9, 317
67. Kumar Sharma, R. *et al.* (2019) Complex DNA knots detected with a nanopore sensor. *Nat. Commun.* 10, 4473
68. Witz, G. *et al.* (2011) Tightening of DNA knots by supercoiling facilitates their unknotting by type II DNA topoisomerases. *Proc. Natl. Acad. Sci. U. S. A.* 108, 3608–3611
69. van Berkum, N.L. *et al.* (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* 39, e1869
70. Dixon, J.R. *et al.* (2016) Chromatin domains: the unit of chromosome organization. *Mol. Cell* 62, 668–680
71. Winfree, E. *et al.* (1998) Design and self-assembly of two-dimensional DNA crystals. *Nature* 394, 539–544
72. Rothemund, P.W.K. *et al.* (2004) Design and characterization of programmable DNA nanotubes. *J. Am. Chem. Soc.* 126, 16344–16352
73. Douglas, S.M. *et al.* (2009) Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* 459, 414–418
74. Seeman, N.C. (2010) Nanomaterials based on DNA. *Annu. Rev. Biochem.* 79, 65–87
75. Rothemund, P.W.K. (2006) Folding DNA to create nanoscale shapes and patterns. *Nature* 440, 297–302
76. Kočar, V. *et al.* (2016) Design principles for rapid folding of knotted DNA nanostructures. *Nat. Commun.* 7, 10803
77. Han, D. *et al.* (2017) Single-stranded DNA and RNA origami. *Science* 358, eaao2648
78. Lin, Y.-R. *et al.* (2015) Control over overall shape and size in *de novo* designed proteins. *Proc. Natl. Acad. Sci. U. S. A.* 112, E5478–E5485
79. Baker, D. (2006) Prediction and design of macromolecular structures and interactions. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 361, 459–463
80. Ljubetič, A. *et al.* (2017) Design of coiled-coil protein-origami cages that self-assemble *in vitro* and *in vivo*. *Nat. Biotechnol.* 35, 1094–1101
81. Donate, L.E. *et al.* (1996) Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci.* 5, 2600–2616
82. Denis, M. and Goldup, S.M. (2017) The active template approach to interlocked molecules. *Nat. Rev. Chem.* 1, 0061
83. Voorhaar, L. and Hoogenboom, R. (2016) Supramolecular polymer networks: hydrogels and bulk materials. *Chem. Soc. Rev.* 45, 4013–4031
84. Zhang, L. *et al.* (2019) Effects of knot tightness at the molecular level. *Proc. Natl. Acad. Sci. U. S. A.* 116, 2452–2457
85. Segawa, Y. *et al.* (2019) Topological molecular nanocarbons: all-benzene catenane and trefoil knot. *Science* 365, 272–276
86. Barran, P.E. *et al.* (2011) Active-metal template synthesis of a molecular trefoil knot. *Angew. Chem.* 123, 12488–12492
87. Ponnuswamy, N. *et al.* (2014) Homochiral and meso figure eight knots and a Solomon link. *J. Am. Chem. Soc.* 136, 8243–8251
88. Ayme, J.-F. *et al.* (2012) A synthetic molecular pentafoil knot. *Nat. Chem.* 4, 15–20
89. Danon, J.J. *et al.* (2017) Braiding a molecular knot with eight crossings. *Science* 355, 159–162
90. Nikoofard, N. and Mashaghi, A. (2016) Topology sorting and characterization of folded polymers using nano-pores. *Nanoscale* 8, 4643–4649
91. Cang, Z. *et al.* (2018) Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* 14, e1005929