Topological decoding of biomolecular fold complexity

Scalvini, B.

CHAPTER 7:

CONCLUSION AND OUTLOOK

In this thesis, we have covered substantial advancements in the field of contact-based topology, specifically in the development of the Circuit Topology (CT) framework for the analysis of real biopolymers. This thesis contributed to the creation of computational pipelines [1] for CT characterization of proteins (including IDPs) and genome structures, therefore creating the first real world applications of what was previously only a conceptual model. In Chapter 2 we demonstrated the predictive power of native circuit topology concerning protein kinetics. Moreover, we unveiled a fundamental principle for fast protein folding, which we called *zipper effect*: how an abundance of parallel and cross relations in the native conformation correlates with high folding rates. The simple trends between topology and kinetics highlighted in Chapter 2 not only shed a light on the folding process as it happens in nature, but also provide a topological prescription for synthetic polymers with the desired biophysical properties [2][3]. Synthetic polymers have witnessed in popularity for the production of biopharmaceutical constructs, ranging from drug carriers to polymer-drug conjugates and polymeric drugs[4]; however, only few of these new materials have so far entered the market, and the potential of such technology remains mostly untapped.

The power of the CT representation is especially apparent when studying the folding process from a dynamic point of view. Spontaneous folding of proteins occurs in a time span ranging from milliseconds to seconds, in spite of the astronomical number of possible configurations, as highlighted by Levinthal with his famous paradox in 1969. A smart way to visualize real folding pathways would undeniably provide much insight in discerning among the various models and views of the folding pathways that have been suggested ever since [5]. In this context, the graphical representation of protein conformational evolution on the 3D topological landscape shown in Chapter 3 represents a technological advancement in itself. We demonstrate the potential of such framework for the identification of transient states in Intrinsically Disordered Protein (IDP) dynamics, an endeavour notoriously challenging in light of the flat energy landscape characterizing this protein class. However, such representation is general and could be applied to any polymer undergoing configurational evolution. It could also be applied for visualization of folding upon ligand binding, a phenomenon not uncommon for IDPs [6] and potentially crucial for drug targeting of IDP regions [7]. Moreover, our novel definition of topological similarity for IDPs promises to be key in further homology and classification studies, as it does not depend, like traditional methods, on either sequence or secondary structure of the protein.

In the cell, topological transitions of biopolymers are often a collaborative process, with a variety of ligands, proteins and molecular chaperones interacting and posing constraints to the folding process. Such complexity is often overlooked in

force spectroscopy assays, where specific interactions between protein client and chaperone systems are studied in isolation. In Chapter 4 we proposed a paradigm shift, utilizing a widely studied model protein (MBP) as a topological sensor to test the structural impact of the cytosolic interactome. We demonstrate how such cumulative effect promotes stabilization against forced unfolding. This type of experiments might constitute a fundamental step towards the validation of optical tweezers experiments as assays capable of mimicking in vivo biological conditions. Moreover, the concept of a topological sensor used as proxy biomarker for chaperone activity could prove pertinent for clinically relevant cellular states involving modifications of the cellular chaperone content [8]. The introduction of formalism and representation methods borrowed from circuit topology – such as the circuit diagram – to analyse folding pathways enabled us to draw conclusions from otherwise noisy and heterogeneous data. As such, CT has the potential to become a key intermediate step to bridge force spectroscopy data and their structural interpretation.

A smart topological encoding is particularly crucial when dealing with very extended, complex folded polymers, where dimensionality reduction is necessary to extract meaningful information and patterns. As such CT found a very natural application in the characterization of genome architecture, a field where an abundance of contact-based arrangement data (Hi-C) still requires appropriate and unified analysis standards. Key findings in this respect were the identification and description of highly conserved conformational patterns in single cell configurations and a multi-scale heterogeneity pipeline able to quantify single cell heterogeneity in genome structure (Chapter 5). The latter is particularly relevant if we consider the role that gene annotation methods such as Hi-C can have in the identification of harmful interactions and variants [9]. Further developments in our CT based pipeline, including its extension to population Hi-C, could support the understanding of epigenetic factors leading to genetic disorders and assist biologists and clinicians in their assessment. Obvious next steps in this line of research involve topological mapping of specific chromosome features, such as Topologically Associated Domains and chromatin loops, at different levels of resolution. Moreover, the topological fingerprint of each stage of the cell cycle should be assessed, in order to provide a baseline for discerning health and disease conditions.

The variety of systems explored in this thesis is by itself a testimony of the flexibility of the framework. Ideally, any linear chain where direction and intra-chain contacts can be formally defined could be object of CT analysis. In Chapter 6, we applied our formalism to a system outside the realm of biopolymers, by treating texts from true and fake news articles as sentence-chains where contacts

are identified by semantic similarity. By this simple device, we managed to detect statistical differences in the semantic similarity arrangement in the two datasets, highlighting what could be a characteristic feature of fake news communication. To test this hypothesis, future work should consider feeding topological parameters encoded as linguistic features into classification algorithms for fake news detection. This proof-of-concept work demonstrates the wide range of applicability of Circuit Topology to issues of societal relevance at large. There are, anyways, challenges and opportunities in the development of the framework which future practitioner should take into account. We summarize some of these key points below:

- **Back mapping of topological findings onto biopolymer structures.** In this thesis, we explore ways to encode structural features of 3D polymers into the topological space. The new representations thus created allow for the identification of patterns and conformation, as it is particularly evident in the case of IDPs. However, projection of the observed findings back to the 3D space – both in terms of structural coordinates and biological significance – is not always obvious. As such, one of the main challenges and clear direction for further development of the framework is the formalization of an appropriate transformation for translating topological parameter in terms of local structural properties. For what concern genome topology studies, we either segment the structure in order to gain local information (with limited resolution) or sum all topological relations in which one contact site is involved in order to get a feeling of its impact over the rest of the structure (Chapter 5). A similar approach was employed by Woodard et al.[10], with the definition of local Circuit Topology, which proved to have the potential to predict the pathogenicity of missense mutations.

  These simple devices proved effective in the cases at hand. However, they might not be as effective for other types of structures, such as smaller proteins and peptides, and they anyway lack extensive theoretical proof and formalization. Topological parameters as formalized by CT are by definition non local, as they reflect the need to reduce the complexity of the configurational search problem by removing 3D coordinates; also, they are defined by pairs of contacts, which are in turn formed by two contact sites. Therefore, 4 different coordinates are associated to one topological relation, in case of a static structure. It is unclear how the relation should be accounted for or weighted among these coordinates. It is easy to see how there is no trivial choice of backwards transformation from the topological to the physical space, but a formalization of possible alternatives would be beneficial for specific applications. The possibility to localize topological para-

meters would prove especially crucial in the case of IDPs, where identifying structural motifs for drug binding is paramount for any pharmaceutical application [11].

- **Accounting for backbone contribution.** While the theoretical formalism of CT has in recent years been extended to include the contribution of backbone crossings and entanglement [12] (the so-called soft contacts), this extension is yet to be translated into software applications for real biopolymers. Therefore, the applications discussed in this thesis only take into account hard contacts, which can be considered glued or static in the configuration under study. While knots are extremely rare in proteins, the influence of various types of backbone entanglement over the protein's kinetic properties is well documented [13][14]. Therefore, it is advisable to extend the current state of available pipelines to include soft contacts, in order to provide a complete topological description.
- **Investigate conceptual links and applications of CT relations in networks.** CT shares with another topology framework, network topology, its focus on contacts and connectivity. We exploited this conceptual affinity in Chapter 5, where we coupled CT analysis with a network parameter such as the average clustering coefficient of the network representation of genome architecture. Such combination of frameworks can provide a multi-faceted representation; while network topology focuses on statistical properties, CT can provide an unambiguous topological description of the individual chain, thus allowing to discern between polymers with similar connectivity. Small world networks are a particular class of networks characterized by small average path length and high clustering coefficient[15]. Such characteristics make them particularly suitable for modelling protein structure and connectivity; leveraging this representation, it has been possible to identify fundamental structural features such as the existence of key residues involved in protein folding [16]–[18]. These residues are characterized by a high number of contacts with other residues in the transient state, forming thus the folding nucleus of the native configuration. If we were to represent such phenomenon in CT terms, we would find an increased number of concerted relations (CS, CP) surrounding these key residues, as many contacts include the key residue as one of the two contact sites. It is thus easy to interpret in this light the results obtained in Chapter 5, where we saw how average clustering coefficient values solely depended on concerted relations. The formation of clusters – those same clusters that confer the network its 'small world' characteristics – is driven by these key contact sites forming a higher number of connections with respect to their neigh-

bours, thus increasing CP and CS relations. Moreover, the small average path length of the small world network hints at the presence of non-local connections, which in turn could be related to a higher percentage of entangled relations: these can indeed be created by non-local, longer-range contacts enveloping shorter-range loops. While small world network analysis proved very successful in the identification of key residues[16]–[18] and hot spots for protein-protein interaction [19][20], some questions remain open. For example, such analysis is plagued by false positives[17], residues that present (often in the native state) higher connectivity but are not crucial for the folding process at the transient state level. A CT description might be able to distinguish between true and false positives, by looking at, for example local CP/CS ratio, local topology of other structural elements, and most importantly, circuit analysis. Circuit analysis, exemplified in Chapter 2, allows for deconstruction of independent structural subunits of the protein; an interesting future application, in this sense, could test possible correlations between circuit and key residue distribution.

- **Applying the CT framework to multi-chain systems.** To deal with systems composed of multiple chains, we need to extend classic CT representation to include intra-chain contacts. Such new generalization of the framework involves the definition of additional configurations (loop, tandem and independent) on top of the three topological relations [21], and is not the object of this thesis. We can, however, speculate about its future applications for topological characterization of macromolecular complexes and condensates. Macromolecular condensates, also called membraneless organelles, have been in recent times object of intense exploration, due to their dynamic physical nature [22]–[25], biological function [26][27], pharmaceutical potential [28][29] and the still unclear assembly mechanism, involving liquid-liquid phase separation [30]–[35]. Such systems can be formed by a variety of polymer-like molecules, such as IDPs, proteins with disordered regions (IDRs), DNA and RNA. Condensate components are generally classified as scaffolds, if they are responsible for phase formation, or clients, when they are recruited by interaction with the scaffold [36]. However, the same components have been proven to sometimes play both roles in complex assembly [37][38], indicating that the mechanism might be non-specific and topological in nature. Multi-chain circuit topology might shed a light on the formation of such condensates. Moreover, these systems are characterized by a hierarchy of forces: strong interactions ensure structure, while weaker interaction give the condensate its dynamic properties [24]. One could consider uncoupling the role of strong and weak interactions and study their topological evolution and multi timesca-

le dynamics, with a multi-chain generalisation of the techniques displayed in Chapter 3. Identifying topological features in biomolecular condensates would be a fundamental step in understanding their assembly mechanisms and function.

- **Exploring the potential of CT for ML applications.** The rise of machine learning techniques in recent years has brought unprecedented progress to the field of protein structure prediction, such as the milestone algorithm Alphafold2. However powerful these models could be, they remain limited by quality and quantity of data and data representations. In this context, the CT formalism could prove to be advantageous, in view of its ability to highlight key conformational features in the topology space. However, the specifics of this type of application are still to be explored. Taken in its simplest form, the number of P, S and X relations does not constitute an ideal data basis for many ML algorithms, because the three variables are often found to strongly correlate. In this thesis we have overcome the issue by either excluding one of the topological relations as redundant (linear regression in Chapter 2). Either way, a formal investigation should be undertaken, to explore what transformations of the CT parameters could be most efficient and meaningful, depending on the chosen model and scientific problem at hand. Genome topology represents possibly the most promising application for CT-based predictive analysis. The wealth and depth of population Hi-C interaction matrices available would allow for training of bigger neural networks models, with the aim of classifying potentially harmful variations in genome arrangement. Another key application would involve performing pattern recognition over topology matrices. The topology matrix represents the topological fingerprint of the biopolymer under analysis, the significance of which we have only started to uncover in this thesis. Systematic and extensive feature-extraction performed on large protein datasets would certainly bring to light many conformational motifs of biophysical relevance.

REFERENCES

[1] D. Moes, E. Banijamali, V. Sheikhhassani, B. Scalvini, J. Woodard, and A. Mashaghi, "ProteinCT: An implementation of the protein circuit topology framework," MethodsX, vol. 9, Jan. 2022, doi: 10.1016/j.mex.2022.101861.

[2] X. W. Wang and W. bin Zhang, "Chemical Topology and Complexity of Protein Architectures," Trends in Biochemical Sciences, vol. 43, no. 10. Elsevier Ltd, pp. 806–817, Oct. 01, 2018. doi: 10.1016/j.tibs.2018.07.001.

[3] A. Ljubetič et al., "Design of coiled-coil protein-origami cages that self-assemble in vi-

tro and in vivo," Nat Biotechnol, vol. 35, no. 11, pp. 1094–1101, Nov. 2017, doi: 10.1038/nbt.3994.

[4] C. Englert et al., "Pharmapolymers in the 21st century: Synthetic polymers in drug delivery applications," Progress in Polymer Science, vol. 87. Elsevier Ltd, pp. 107–164, Dec. 01, 2018. doi: 10.1016/j.progpolymsci.2018.07.005.

[5] S. W. Englander and L. Mayne, "The nature of protein folding pathways," Proceedings of the National Academy of Sciences of the United States of America, vol. 111, no. 45. National Academy of Sciences, pp. 15873–15880, Nov. 11, 2014. doi: 10.1073/pnas.1411798111.

[6] P. Robustelli, S. Piana, D. E. Shaw, and D. E. Shaw, "Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein," J Am Chem Soc, vol. 142, no. 25, pp. 11092–11101, Jun. 2020, doi: 10.1021/jacs.0c03217.

[7] C. Y. C. Chen and W. I. Tou, "How to design a drug for the disordered proteins?," Drug Discov Today, vol. 18, no. 19–20, pp. 910–915, 2013, doi: 10.1016/j.drudis.2013.04.008.

[8] L. Goltermann, M. v. Sarusie, and T. Bentin, "Chaperonin GroEL/GroES Over-Expression Promotes Aminoglycoside Resistance and Reduces Drug Susceptibilities in Escherichia coli Following Exposure to Sublethal Aminoglycoside Doses," Front Microbiol, vol. 6, Jan. 2016, doi: 10.3389/fmicb.2015.01572.

[9] C. A. Steward, A. P. J. Parker, B. A. Minassian, S. M. Sisodiya, A. Frankish, and J. Harrow, "Genome annotation for clinical genomic diagnostics: Strengths and weaknesses," Genome Medicine, vol. 9, no. 1. BioMed Central Ltd., May 30, 2017. doi: 10.1186/s13073-017-0441-1.

[10] J. Woodard, S. Iqbal, and A. Mashaghi, "Circuit topology predicts pathogenicity of missense mutations," Proteins: Structure, Function and Bioinformatics, vol. 90, no. 9, pp. 1634–1644, Sep. 2022, doi: 10.1002/prot.26342.

[11] Q. H. Chen and V. v. Krishnan, "Identification of ligand binding sites in intrinsically disordered proteins with a differential binding score," Sci Rep, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-00869-4.

[12] A. Golovnev and A. Mashaghi, "Generalized Circuit Topology of Folded Linear Chains," iScience, vol. 23, no. 9, p. 101492, 2020, doi: 10.1016/j.isci.2020.101492.

[13] E. Panagiotou and K. W. Plaxco, "A topological study of protein folding kinetics," ArXiv, pp. 1–13, 2018, doi: 10.1090/conm/746/15010.

[14] M. Baiesi, E. Orlandini, F. Seno, and A. Trovato, "Exploring the correlation between the folding rates of proteins and the entanglement of their native states," ArXiv, 2017.

[15] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature, vol. 393, no. 6684, pp. 440–442, Jun. 1998, doi: 10.1038/30918.

[16] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, "Three key residues form a critical contact network in a protein folding transition state," Nature, vol. 409, no. 6820, pp. 641–645, Feb. 2001, doi: 10.1038/35054591.

[17] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, "Small-world view of the amino acids that play a key role in protein folding," Phys Rev E, vol. 65, no. 6, p. 061910, Jun. 2002, doi: 10.1103/PhysRevE.65.061910.

[18] A. R. Atilgan, P. Akan, and C. Baysal, "Small-World Communication of Residues and Significance for Protein Dynamics," 2004.

[19] A. del Sol, H. Fujihashi, and P. O'Meara, "Topology of small-world networks of protein-protein complex structures," Bioinformatics, vol. 21, no. 8, pp. 1311–1315, Apr. 2005, doi: 10.1093/bioinformatics/bti167.

[20] A. Del Sol and P. O'Meara, "Small-world network approach to identify key residues in protein-protein interaction," Proteins: Structure, Function and Genetics, vol. 58, no. 3, pp. 672–682, Feb. 2005, doi: 10.1002/prot.20348.

[21] M. Heidari, D. Moes, O. Schullian, B. Scalvini, and A. Mashaghi, "A topology framework for macromolecular complexes and condensates," Nano Res, vol. 15, no. 11, pp. 9809–9817, Nov. 2022, doi: 10.1007/s12274-022-4355-x.

[22] C. Yu, Y. Lang, C. Hou, E. Yang, X. Ren, and T. Li, "Distinctive Network Topology of Phase-Separated Proteins in Human Interactome," J Mol Biol, vol. 434, no. 1, Jan. 2022, doi: 10.1016/j.jmb.2021.167292.

[23] D. M. Mitrea et al., "Methods for Physical Characterization of Phase-Separated Bodies and Membrane-less Organelles," Journal of Molecular Biology, vol. 430, no. 23. Academic Press, pp. 4773–4805, Nov. 02, 2018. doi: 10.1016/j.jmb.2018.07.006.

[24] J. D. Schmit, M. Feric, and M. Dundr, "How Hierarchical Interactions Make Membraneless Organelles Tick Like Clockwork," Trends in Biochemical Sciences, vol. 46, no. 7. Elsevier Ltd, pp. 525–534, Jul. 01, 2021. doi: 10.1016/j.tibs.2020.12.011.

[25] L. Jawerth et al., "Protein condensates as aging Maxwell fluids," Science (1979), vol. 370, no. 6522, pp. 1317–1323, 2020, doi: 10.1126/science.aaw4951.

[26] A. Boija, I. A. Klein, and R. A. Young, "Biomolecular Condensates and Cancer," Cancer Cell, vol. 39, no. 2, pp. 174–192, 2021, doi: 10.1016/j.ccell.2020.12.003.

[27] B. R. Sabari, "Biomolecular condensates in the nucleus," 2020.

[28] I. A. Klein et al., "Partitioning of cancer therapeutics in nuclear condensates," Science (1979), vol. 368, no. 6497, pp. 1386–1392, 2020, doi: 10.1126/science.aaz4427.

[29] M. Biesaga, M. Frigolé-vivas, and X. Salvatella, "ScienceDirect Intrinsically disordered proteins and biomolecular condensates as drug targets," Curr Opin Chem Biol, vol. 62, pp. 90–100, 2021, doi: 10.1016/j.cbpa.2021.02.009.

[30] J. D. Schmit, J. J. Bouchard, E. W. Martin, and T. Mittag, "Protein Network Structure Enables Switching between Liquid and Gel States," J Am Chem Soc, vol. 142, no. 2, pp. 874–883, 2020, doi: 10.1021/jacs.9b10066.

[31] C. P. Brangwynne et al., "Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation," Science (1979), vol. 324, no. 5935, pp. 1729–1732, Jun. 2009, doi: 10.1126/science.1172046.

[32] C. P. Brangwynne, T. J. Mitchison, and A. A. Hyman, "Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes," Proceedings of the National Academy of Sciences, vol. 108, no. 11, pp. 4334–4339, Mar. 2011, doi: 10.1073/pnas.1017150108.

[33] S. Elbaum-Garfinkle et al., "The disordered P granule protein LAF-1 drives phase separa-

tion into droplets with tunable viscosity and dynamics," Proceedings of the National Academy of Sciences, vol. 112, no. 23, pp. 7189–7194, Jun. 2015, doi: 10.1073/pnas.1504822112.

[34] A. Molliex et al., "Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization," Cell, vol. 163, no. 1, pp. 123–133, Sep. 2015, doi: 10.1016/j.cell.2015.09.015.

[35] Y. Lin, D. S. W. Protter, M. K. Rosen, and R. Parker, "Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins," Mol Cell, vol. 60, no. 2, pp. 208–219, Oct. 2015, doi: 10.1016/j.molcel.2015.08.018.

[36] S. F. Banani et al., "Compositional Control of Phase-Separated Cellular Bodies," Cell, vol. 166, no. 3, pp. 651–663, Jul. 2016, doi: 10.1016/j.cell.2016.06.010.

[37] J. A. Ditlev, L. B. Case, and M. K. Rosen, "Who's In and Who's Out—Compositional Control of Biomolecular Condensates," J Mol Biol, vol. 430, no. 23, pp. 4666–4684, Nov. 2018, doi: 10.1016/j.jmb.2018.08.003.

[38] J. A. Riback et al., "Composition-dependent thermodynamics of intracellular phase separation," Nature, vol. 581, no. 7807, pp. 209–214, May 2020, doi: 10.1038/s41586-020-2256-2.