UNIVERSITY OF
LATVIA

Summary of
Doctoral Thesis

Gatis Melkus

GRAPH-BASED METHODS FOR MODELING BIOMOLECULAR NETWORKS

Riga 2025

UNIVERSITY OF
LATVIA

FACULTY OF SCIENCE AND TECHNOLOGY

Gatis Melkus

GRAPH-BASED METHODS FOR MODELING BIOMOLECULAR NETWORKS

SUMMARY OF THE DOCTORAL THESIS

Submitted for the degree of Doctor of Engineering and Technology
Field of Computer Science and Informatics
Subfield of Bioinformatics

Riga 2025

This doctoral thesis was carried out: at the Chair of Computer Science, Faculty of Science and Technology, University of Latvia from 2019 to 2025.

INVESTING IN YOUR FUTURE

Eiropas Sociālā fonda projekts "LU doktorantūras kapacitātes stiprināšana jaunā doktorantūras modeļa ietvarā" Nr. 8.2.2.0/20/I/006

The thesis contains the introduction, 3 main sections (including five scientific publications reproduced in full), conclusions and no appendices.

Form of the thesis: collection of research papers in computer science (mathematical foundations of computer science).

Supervisor: Dr. sc. comp. **Juris Vīksna**, prof.

Reviewers:

1. Dr. sc. comp. **Ģirts Karnītis**, prof., University of Latvia;
2. Dr. sc. ing. **Egils Stalidzāns**, prof., Rīga Stradiņš University;
3. Dr. sc. comp. **Alvis Brāzma**, European Bioinformatics Institute (EMBL-EBI), Scientist Emeritus.

This thesis will be defended at the public session of the Doctoral Committee of Computer Science and Informatics and of Electrical Engineering, Electronics and Communication Technologies, University of Latvia on October 24, 2025.

The thesis is available at the Library of the University of Latvia, Kalpaka blvd. 4.

Chairman of the Doctoral Committee Dr. sc. comp. **Guntis Bārzdiņš**, prof.

Secretary of the Doctoral Committee **Sintija Siliņa**

ISBN 978-9934-36-422-8
ISBN 978-9934-36-423-5 (PDF)

ABSTRACT

The analysis and interpretation of biomolecular data touches on many general issues in biology as well as the sciences as a whole, most notably the issues of big data and sustainable research practices. Among these issues is the continually increasing complexity of biological data sets, including both the scale of individual data sets, the variety of methodologies and the escalating difficulty of reliable and robust comparison when attempting to integrate disparate data. This thesis employs the familiar paradigm of graph theory and network biology to generate novel methods of analyzing biomolecular data. It is formatted as a collection of five published papers by the author, one of them in the established field of gene regulatory network comparison and modeling, while the remaining four turn to the emerging field of chromatin interaction network models based on high-throughput chromatin conformation capture data. Alongside these is a brief review of key concepts in network biology, gene regulatory network modeling and chromatin conformation capture in order to give proper context to the work accomplished over the course of the author's studies. The results obtained over the course of the scientific work undertaken show plentiful potential in chromatin interaction network models as a way of comparing and parsing chromatin interaction data. Also demonstrated are several novel graph-based methods for both gene regulatory network analysis and development as well as a framework for analyzing processed chromatin conformation capture data and matching up network topology to other biological data for validation purposes.

Keywords: network biology, gene regulatory network, Hi-C, network topology, integrative biology.

TABLE OF CONTENTS

INTRODUCTION

Relevance of the thesis

The introduction of graph-based methodologies to molecular biology at the turn of the century has led to the rise of the field of network biology. Network biology, as distinct from systems biology which focuses on functional mechanisms, is concerned with the concept of interactions and networks formed out of such interactions. These networks are generally represented as graphs which can be used to model a large number of biomolecule interactions such as protein-protein interactions, genetic interactions, gene regulation events and chromatin architecture. A benefit of this approach is the ability of researchers to use graph theory concepts in representing biological processes and consequently employ approaches from topology and related fields to construct elaborate models of biological function that may produce a more in-depth understanding of how biological interactions might add up to a larger whole (Barabási and Oltvai 2004; Sorrells and Johnson 2015).

Owing to the large and specialized field of molecular biology producing a wide variety of datasets, there is a wealth of publicly available data that may be used to develop, test and refine network models of biological phenomena. In addition to sequence databases and repositories of full experiments, there are additional resources that provide examples of networks and supplementary data to perform a wide range of analyses and validate findings from network biology models. However, while there has been substantial work in bridging gaps in knowledge using network models, the newness of network biology as a field and the development of ever-newer data gathering methods mean that graph theory methods as well as datasets are heterogeneous and difficult to conclusively evaluate for optimal performance. This is the case for gene regulatory networks (Sorrells and Johnson 2015), but especially so for fields such as chromatin spatial organization where the methods of integrating and cross-validating data in an efficient manner remain an open question and the development of new methods may well allow for deeper understanding of the structures being studied (Pancaldi 2021; 2023).

Due to an overabundance of data and the relative paucity of steady conclusions, reliable methodologies for mapping graph topology and similar concepts to biological function remain relevant, especially with increasing opportunities for integrative data analysis with the advent of multi-omics datasets alongside exhaustive single-cell assays (Vandereyken et al. 2023) that

may make it simpler to derive reliable conclusions about how, if at all, the basic rules dictating the topology of biomolecule interactions may work.

Research goals and objectives

The goal of the research undertaken as part of this thesis was to establish novel methods of graph-based analysis for biomolecular networks that utilize graph topology as their primary metric. This can be subdivided into the following key objectives:

1. To extend existing topological methods used for studying network motifs and graphlets to survey topological similarity between gene pairs in gene regulatory networks.
2. To develop new topological approaches to studying chromatin interaction networks through the use of connected components and derived concepts.
3. To discover functionally relevant subgraphs such as gene regulatory modules within larger chromatin interaction networks via topological means.
4. To use graph-based methods in conjunction with additional biomolecular data to validate our findings.
5. To refine and select the most effective topological criterion for isolating functionally relevant subgraphs within a chromatin interaction network.

Primary theses

With the above objectives in mind, the primary hypothesis of the present work is that biological properties can be identified in biomolecular data using purely topological graph-based means (such as the presence of a particular network motif or divergent counts of types of subgraph such as cliques). Working from this hypothesis, the following theses encompass the primary outcomes of this work:

- The bi-fan motif, when generalized as the bi-fan unit, is a modular, easy-to-embed tool for capturing patterns of gene regulatory network symmetry between paralogous gene pairs in various organisms that do not exist in non-paralogous genes.
- High-throughput chromatin conformation capture (Hi-C) data can be topologically characterized by identifying, separating and analyzing connected components in-depth to discover functional modules in them even in the absence of other biological information.
- Through narrower applications of specific topological elements, we can identify functionally relevant subgraphs in chromatin interaction networks that are enriched in features such as active gene expression and epigenetic regulatory marks.

- Graph-based models for chromatin conformation capture are helpful in pulling together and integrating information about combinations of features such as positional gene expression data, chromatin annotation data and others.
- Of the graph-theoretical features most promising to examine, the foremost are cliques, which map not only to hotspots of regulation but also constitute formations of chromatin that are most valuable to characterize.

Overview of literature

The scope of this thesis encompasses both the study of network motifs, a phenomenon historically observed in gene regulatory networks, and the similar but distinct field of chromatin interaction networks. Both fields involve topological analyses in their essential methods, though the treatment of topology in gene regulatory networks is more extensive.

While network motifs were not novel to biology as a whole owing to their historic use in ecological models, they began to be used to describe interaction networks with the seminal research of the Alon lab at the turn of the 21st century (Babu et al. 2004; Alon 2007; Stone, Simberloff, and Artzy-Randrup 2019). They defined network motifs as subgraphs within a larger network of transcription factors and their regulatory targets (genes, which could also be transcription factors themselves) that were overrepresented in that network compared to what would be expected in a mathematically modeled network with a similar overall topology. Many of these were likewise assigned putative functions in the biochemical signaling network, such as delaying or modulating responses to particular stimuli. Since that time a substantial amount of work has been invested in developing and testing this assertion, and though similar network biology theories such as the prevalence of scale-free networks in nature have since been strongly disputed (Lima-Mendez and van Helden 2009; Broido and Clauset 2019), network motif-like concepts continue to see use in concepts such as graphlet-based analyses (Pržulj 2007; Sarajlić et al. 2016).

The particular interest of this thesis is the existence of motifs known as bi-fan motifs (Ward and Thornton 2007) which, unlike motifs such as feed-forward loops which are implicated in signal modulation and similar functions, have been suggested to be a result of gene duplication events across a species' evolutionary history. Since there are many traceable gene duplication events that have been historically proven, particularly in model organisms such as the baker's yeast *Saccharomyces cerevisiae* which has undergone at least one whole-genome duplication with a full list of genes annotated and listed in publicly available resources alongside a reasonably accurate gene regulatory network (Byrne and Wolfe 2005; Teixeira et al. 2023), our objective here was to investigate in more

detail the possibility of a network motif having a direct relationship to a gene pair's phylogenetic history.

By comparison to the more indirectly observed information depicted in a gene regulatory network, chromatin interaction networks directly depict the spatial organization of chromatin. Chromatin, a complex of proteins, non-coding RNAs and DNA, is both the result and the subject of substantial genomic regulation, and the exact conformation of DNA within the nucleus of a eukaryotic cell is both vitally important and incompletely understood. Much of the currently known details about how DNA is organized to facilitate RNA transcription more narrowly and gene regulation more broadly is the result of chromatin conformation capture experiments. These experiments, most notably high-throughput chromatin conformation capture or Hi-C (Dekker et al. 2002; Lieberman-Aiden et al. 2009), produce "contact maps" of chromatin interactions that represent regions of the genome coming into close contact, which are often indicators of purposeful chromatin architecture such as topologically associating domains, promoter-enhancer interactions, chromatin compartments and similar phenomena. Many of these concepts, in fact, were originally defined statistically through the application of methods such as PCA performed on Hi-C data and later observed through other methods (Lajoie, Dekker, and Kaplan 2015; Pancaldi 2023). Computational methods are not only vital to interpreting Hi-C data, they are also critical to identifying features in the data that may be key to discovering novel biological functions and nuances in gene regulation.

Overview of methods

All of the biomolecular data sourced for the papers published was obtained from previously completed scientific studies and public repositories. The wide availability of data in the field of computational biology means that, while provenance of an individual data set is still a concern, there was ultimately no shortage of either gene regulatory network data that could be sourced from public repositories such as YEASTRACT and RegulonDB (Teixeira et al. 2023; Tierrafría et al. 2022) or chromatin interaction network data sourced from individual experiments (Javierre et al. 2016; Jung et al. 2019; Kim et al. 2021) supplemented with publicly available genome annotations from accredited databases such as Ensembl and ENCODE (Harrison et al. 2024; Abascal et al. 2020) to test the prevalence of biologically significant properties as well as expression data from atlases such as the Genotype Tissue Expression (GTEx) or FANTOM5 atlases (Lonsdale et al. 2013; Noguchi et al. 2017) to obtain quantifiable differences in gene products depending on regulation.

With this data in hand, we generally implemented a pipeline of rendering our original dataset – either a set of gene regulatory events or chromatin interactions – into a graph. In these graphs, genomic loci or genes generally

constitute vertices while interactions between them (should they exist) constitute edges. In each case they are generated directly from a given data set, usually by filtering the interactions by a criterion of statistical reliability such as p-value included in the data and assessed from the original experiments (for the exact criteria used in each case, consult the individual papers). Then, using the graphs generated, we proceeded to find topological features of interest in them in the form of various kinds of subgraphs. These include network motifs (Publication I), connected components (Publications II–IV) and cliques (Publication V), which can then be analyzed and compared for further variation within that category. For the most part we employed counting metrics (normalized counts of topological features) for assessing topological difference, with some additional refinement and variations as the work goes on. There are definite algorithmic challenges and refinements employed as part of this process, but these are not the focus of the work and will not be addressed in detail in the thesis itself, which is focused more on the viability of topological metrics as a computational biology technique. Furthermore, a related corpus of visualization and randomization techniques was published separately from the main series of publications in this thesis – interested readers are encouraged to consult these for details (Sizovs, Silina, et al. 2024; Sizovs, Melkus, et al. 2024).

Using bespoke topological measurement approaches of the kinds described above, we developed and tested several approaches of locating interesting features via network topology, which was then further tested by testing independently gathered biological data such as gene expression in conjunction with our categories of interest. These were tested with appropriate statistical analysis methods for a given data set (such as non-parametric analysis of variance) to determine the statistical significance of differences between our selected topological features and known biological features in the organism being studied. The higher amount of biological features observed within our category of interest compared to the data set as a whole was considered evidence of "enrichment", i.e., that the topological features in question are correlated with the biological features in some way. This was extended in numerous ways across the publications covered, and more of this will be covered in the main chapters of the thesis summary.

Overview of results

Owing to the related but distinct graph-based modeling approaches used for different publications in the milieu covered by the present work, the chapters of this thesis are divided by publication. The five publications these results appeared in are reproduced in full.

Our work on gene regulatory networks is covered by **Publication I**, where we discuss the application of graph-based modeling to gene regulatory networks

and make use of graphlets to analyze symmetrical motifs in paralogous gene pairs. We define the concept of the bi-fan unit as well as introduce the idea of complex motifs that include symmetrical positions that may be doubled. We perform this analysis on a series of organisms – baker's yeast, nematode and *Escherichia coli* – and discover that while the latter does not have a noticeable pattern of network symmetry for paralogous genes, the pattern very clearly holds for both yeast and worm networks.

In **Publication II** we turn to chromatin interaction networks and explore the topology of connected components in blood cell promoter capture high-throughput chromatin conformation capture data. We introduce an algorithm for finding network components within a graph formed out of a chromatin interaction contact map and suggest an approach for locating components of biological significance in a set of interrelated blood cell types, demonstrating some initial results in the form of enrichment analyses of our selected components.

In **Publications III and IV**, which overlap substantially in subject matter and continue a course of research from one to the next, we extend our component-based approach to develop a set of topological metrics we compare within our component dataset in order to establish methods of topologically distinguishing different cell types by their mutual relatedness (expressed as "distance" between different pairings of related cells). We use these metrics on our blood cell pcHiC data and successfully find some agreement between some of our proposed distance measures. We also additionally validate the biological significance of our components by expression data analysis using the FANTOM5 promoter-level expression atlas.

In **Publication V** we conclude our analyses of chromatin interaction network data by narrowing our search down to the analysis of cliques of size 3 (or triangles) and aggregations thereof in a several Hi-C datasets, finding significant evidence of enrichment in transcription start sites as well as elevated expression within clique aggregations, particularly when narrowing the comparison RNA polymerase II binding sites. All of this is generally indicative of the presence of "transcription factories", sites of coordinated transcription around closely associated chromatin regions, which our analysis clearly links with localized preponderances of cliques. From here we theorize about potential extensions of the method listed as well as future directions of research.

Approval of the results

The main findings of this thesis are covered by the following publications:

1. Melkus, Gatis, Peteris Rucevskis, Edgars Celms, Kārlis Čerāns, Karlis Freivalds, Paulis Kikusts, Lelde Lace, Mārtiņš Opmanis, Darta Rituma, and Juris Viksna. "Network Motif-Based Analysis of Regulatory Patterns

in Paralogous Gene Pairs." Journal of Bioinformatics and Computational Biology 18, no. 03 (June 18, 2020): 2040008. https://doi.org/10.1142/S0219720020400089. (author contribution: 75%)

2. Viksna, J., G. Melkus, E. Celms, K. Čerāns, K. Freivalds, P. Kikusts, L. Lace, M. Opmanis, D. Rituma, and P. Rucevskis. "Topological Structure Analysis of Chromatin Interaction Networks." BMC Bioinformatics 20 (2019). https://doi.org/10.1186/s12859-019-3237-z. (author contribution: 55%)
3. Lace, Lelde, Gatis Melkus, Peteris Rucevskis, Edgars Celms, Karlis Cerans, Paulis Kikusts, Martiņš Opmanis, Darta Rituma, and Juris Viksna. "Graph-Based Characterisations of Cell Types and Functionally Related Modules in Promoter Capture Hi-C Data." Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019), 78–89, 2019. https://doi.org/10.5220/0007390800780089 (author contribution: 60%)
4. Lace, Lelde, Gatis Melkus, Peteris Rucevskis, Edgars Celms, Kārlis Čerāns, Paulis Ķikusts, Mārtiņš Opmanis, Darta Rituma, and Juris Viksna. "Characteristic Topological Features of Promoter Capture Hi-C Interaction Networks." Communications in Computer and Information Science, vol. 1211, pp. 192–215, 2020. https://doi.org/10.1007/978-3-030-46970-2_10. (author contribution: 65%)
5. Melkus, Gatis, Andrejs Sizovs, Peteris Rucevskis, and Sandra Silina. "Transcriptional Hubs Within Cliques in Ensemble Hi-C Chromatin Interaction Networks." Journal of Computational Biology 31, no. 6 (June 1, 2024): 589–96. https://doi.org/10.1089/cmb.2024.0515. (author contribution: 75%)

In addition to these, several more publications were accepted and released that may be referenced in this thesis but are not directly covered.

- Gatis Melkus, Pēteris Ručevskis, Edgars Celms, Kārlis Čerans, Karlis Freivalds, Paulis Kikusts, Lelde Lāce, Mārtiņš Opmanis, Dārta Rituma, Juris Vīksna. Graph-based network analysis of transcriptional regulation pattern divergence in duplicated yeast gene pairs. ACM International Conference Proceeding Series, 3365954, 2019. https://doi.org/10.1145/3365953.3365954
- Gatis Melkus, Sandra Silina, Andrejs Sizovs, Peteris Rucevskis, Lelde Lace, Edgars Celms, and Juris Viksna. "Clique-Based Topological Characterization of Chromatin Interaction Hubs," 476–86, 2023. https://doi.org/10.1007/978-981-99-7074-2_38.
- Andrejs Sizovs, Gatis Melkus, Peteris Rucevskis, Sandra Silina, Lelde Lace, Edgars Celms, and Juris Viksna. "A Technique for Preserving Network Structure in Randomized Hi-C Data." Journal of Bioinformatics and Computational Biology 22, no. 05 (October 24, 2024). https://doi.org/10.1142/S0219720024400018.

- Andrejs Sizovs, Sandra Silina, Gatis Melkus, Peteris Rucevskis, Lelde Lace, Edgars Celms, and Juris Viksna. "Exploration and Visualization Methods for Chromatin Interaction Data." edited by Wei Peng, Zhipeng Cai, and Pavel Skums, 101–13. Singapore: Springer Nature Singapore, 2024.
- Melkus, Gatis, Karlis Cerans, Karlis Freivalds, Lelde Lace, Darta Zajakina, and Juris Viksna. "Analysis of Dynamics and Stability of Hybrid System Models of Gene Regulatory Networks." In The 12th International Conference on Computational Systems-Biology and Bioinformatics, 1–10. New York, NY, USA: ACM, 2021. https://doi.org/10.1145/3486713.3486727.
- Lace, Lelde, Karlis Cerans, Karlis Freivalds, Gatis Melkus, and Juris Viksna. "Hybrid Gene Regulation Models of Mammalian Circadian Cycles." In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, 130–37. SCITEPRESS – Science and Technology Publications, 2022. https://doi.org/10.5220/0010834400003123.
- Melkus, Gatis, Karlis Cerans, Karlis Freivalds, Lelde Lace, Darta Zajakina, and Juris Viksna. "Behavioral Dynamics of Bacteriophage Gene Regulatory Networks." Journal of Bioinformatics and Computational Biology 20, no. 05 (October 14, 2022). https://doi.org/10.1142/S0219720022500214.
- Viksna, Juris, Karlis Cerans, Lelde Lace, and Gatis Melkus. "Characterizing Behavioural Differentiation in Gene Regulatory Networks with Representation Graphs." NAR Genomics and Bioinformatics 6, no. 3 (July 2, 2024). https://doi.org/10.1093/nargab/lqae102.

In relation to the above, two students involved in the above publications additionally used the results and methods to defend qualifying and bachelor's theses, supervised by the author:

- Sandra Siliņa, 2023. "Vizualizācijas hromatīna interakciju datu analīzei". Qualifying thesis.
- Andrejs Sizovs, 2023. "Sistēmas izstrāde Hi-C datu analīzei". Qualifying thesis.
- Andrejs Sizovs, 2024. "Metodes Hi-C datu randomizācijai, saglabājot to topoloģisko struktūru". Course thesis.
- Andrejs Sizovs, 2024. "Metode tīkla struktūras saglabāšanai randomizētajos hromatīna interakciju datos". Bachelor's thesis.

In addition to the listed publications, the findings in this thesis were presented at several international conferences, either as posters or as oral presentations:

- "Graph-based network analysis of transcriptional regulation pattern divergence in duplicated yeast gene pairs" – CSBio 2019, Nice, France, December 4–7 (oral presentation)
- "Topological features of chromatin interaction networks" – RECOMB 2020, remote (poster)

- "Structural comparison of chromatin interaction networks generated from Hi-C data" – ECCB 2022, Sitges, Spain (poster)
- "Clique-based identification of functional modules in Hi-C graphs" – RECOMB 2023, Istanbul, Türkiye (poster)
- "The utility of cliques in topological characterization of Hi-C data" – ISMB/ECCB 2023, Lyon, France (poster)
- "Clique-based topological characterization of chromatin interaction hubs" – ISBRA 2023, Wroclaw, Poland (oral presentation)
- "Gene Expression Variability Linked to Chromatin Clique Configurations and cis-Regulatory Elements" – ICCBB 2024, Kyoto, Japan (oral presentation)

The main publications covered are included in full in the doctoral thesis as separate chapters, and this summary as well as the thesis itself are both organized around this fact. The thesis is 146 pages long and includes 183 references, excluding references in the text of Publications I–V which are preserved as they originally appeared in the articles in question for clarity.

1. PUBLICATION I – NETWORK MOTIF-BASED ANALYSIS OF PARALOGOUS GENE PAIRS

In this section we discuss the first publication featured in this thesis on the prevalence and topological study of bi-fan motifs as seen in gene regulatory networks. The basic idea of network motifs originates early on in the nascent discipline of network biology, and involves establishing certain topological patterns in gene regulation as overrepresented in a biological network compared to a basic mathematical model (Milo et al. 2002). Several of these such as feedback and feed-forward loops have a large body of literature devoted to them, but in this case we look into the prevalence of the bi-fan motif and its relationship to whole-genome duplications.

Along with more obvious traces of whole genome duplication such as gene synteny (Kellis, Birren, and Lander 2004), the existence of symmetrical bi-fan motifs within gene regulatory networks has been previously linked to gene relatedness (Ward and Thornton 2007). The mechanism at the basis of this is thought to be gene duplication – a freshly duplicated pair of genes obviously should have both the same set of regulators and the same set of regulatory targets. As these genes diverge over the course of evolution, this set of interactions is lost as the sequence of the gene and the structure of the gene product changes accordingly. Hypothetically, this means that interactions in paralogs should ideally show some kind of overlap, not just within the scope of an individual bi-fan motif but also within other graphlets that a gene regulatory network could potentially contain.

To obtain gene regulatory networks suitable for our purposes we employed data from the YEASTRACT dataset of regulators and their targets (Teixeira et al. 2023). From this we constructed a graph where each node is a gene (and its corresponding protein) and each edge is a regulatory interaction. No distinction was drawn between activation and repression here due to longstanding issues with the sign-consistency model (Larsen et al. 2019). To further confirm our results we also assembled the regulatory networks of the bacterium *Escherichia coli* (Tierrafría et al. 2022) as well as the worm *Caenorhabditis elegans* (Reece-Hoyes et al. 2013) from publicly available datasets due to both regulatory networks being reasonably well-known as well as of a manageable size for the construction of graphs. All of these, it should be noted, also have well-established sets of within-species paralogs (genes that originate from duplication within the particular species' history), genes within the same protein family (genes that originate from duplication from before the species came to be) and ohnologs (genes that are preserved from a whole-genome duplication event

sometime in the species' history), which we use as our frame of comparison for how interactions might reflect a given set of genes' evolutionary history.

In order to investigate the basic idea of symmetry within gene regulatory networks we introduce the concept of the bi-fan unit, a simple directed graphlet made up of 3 nodes that contains either a single sink node into which two source nodes direct their path, or two sink nodes with a single source node. This graphlet makes it simpler to reckon the exact size of a bi-fan array, a complex motif that involves an expanding set of source or sink nodes tied to a single pair of nodes they regulate. By integrating the bi-fan unit into a series of classically recognized network motifs, we further develop the idea of symmetry within motifs. Motifs can possess symmetry as a result of particular genes being duplicated, leading to certain positions in these motifs being "twinned", as seen in fig. 1, resulting in complex graphlets with several potentially twinned positions (as opposed to "simple" motifs where these positions are not twinned). This means that bi-fan-based symmetry can be evaluated for a given gene pair by calculating how many of their motifs tend to overlap in this way.

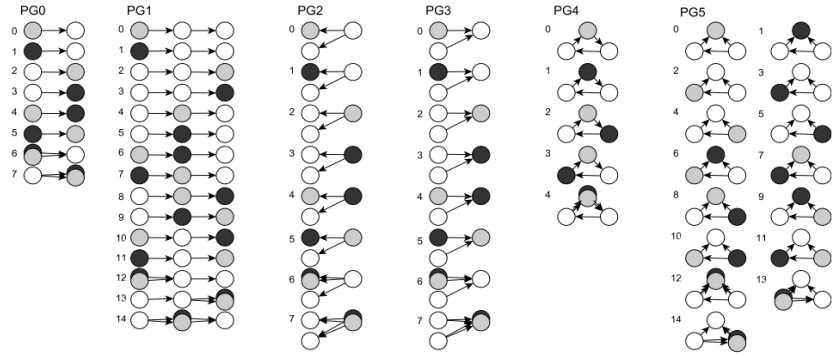

Figure 1. All possible variations of basic graphlets (network motifs) in gene regulatory networks as counted in the design of the study. The gray and black positions indicate the relation of the graphlet to a pair of genes being studied.

In the basic formulation of network motifs, their biological significance is justified by their overrepresentation within a regulatory network compared to what could be explained merely by the connectivity of individual node (Alon 2007). This means that, for example, there should be more feed-forward loops within a yeast regulatory network than can be found within a random network with the same node degrees and overall properties (often modeled as an Erdos-Renyi randomized network, though other options also exist). In our case, we tested whether both simple and complex motifs were significantly depleted

when different proportions of noise (that is, certain proportions of edges were swapped in accordance with the basic properties of the graph).

Results

In analyzing our complex network motif counts, we could readily discern immediate differences in the complex motif count dynamics compared to simple motifs when applying noise to the data. Firstly, the number of complex motifs was an order of magnitude lower than the number of simpler motifs in the yeast data, and certain twinned positions in complex motifs showed substantially larger motif counts than others, most noticeably target gene positions (or Z positions, which have only incoming links and no outgoing ones), which makes sense because genes without regulatory products form the vast majority of the datasets. Secondly, the complex motifs showed substantially higher rates of depletion compared to simple motifs, rapidly declining in number compared to the simple motifs measured in the same dataset. Since the basic concept of the network motif is rooted in these subgraphs being overrepresented in a given biological network, we considered this a solid indication that our complex motifs, rather than simply arising from essential graph properties, did in fact have some kind of biological significance that explained their presence in the network. Thirdly, we noticed none of our paired metrics showed a noteworthy correlation with gene, protein or promoter sequence dissimilarity, which was not strictly surprising given that gene relatedness is often difficult to measure in terms of sequence similarity in general.

By and large, the metrics for particular motifs tended to be closely interlinked, most notably within the confines of a given gene pair where the set of 2-metrics, i.e. the number of motifs featuring the second member of the gene pair, were essentially the bottleneck that determined the number on 12-metrics, i.e. the complex motif featuring both members of the pair. This let us refine our approach and introduce a derived simple measure we called *symmetry,* which is the ratio of 2-metrics to 12-metrics within a given gene pair. This not only helped us evaluate the overall (ostensibly preserved) network topology but also handily scaled the motif counts into a more easily comparable form. We analyzed symmetry between gene pairs (ohnolog, within-species paralog, and other) in yeast, worm and *E. coli* and discovered that in both yeast and worm networks the symmetry between ohnologs as well as within-species paralogs easily and noticeably exceeded that which we observed in proteins from the same family, demonstrating that there is indeed likely to be some preservation in network structure in such paralogs in line with the model outlined in the bi-fan motif concept. The *E. coli* network showed no such thing, however, which could be explained in a number of ways but likely has a lot to do with the relative sparseness, different annotation and less obvious frame of comparison within

a bacterium compared to a eukaryotic organism. This does mean that our results were indeed likely dependent on the specific format of the network, and that future research would be required to properly make general conclusions about the behavior of biological networks.

In any case, this publication was a successful effort in more systematically surveying the bi-fan motif as a measure of gene relatedness, and a good first step in analyzing the prevalence of particular graphlets in networks as a whole. The overall research direction, while promising, ultimately was redirected toward chromatin interaction networks which, as we'll see in the next few chapters, had substantially different considerations structurally as well as topologically compared to the gene regulatory networks examined here.

2. PUBLICATION II – TOPOLOGICAL STRUCTURE ANALYSIS OF CHROMATIN INTERACTION NETWORKS

Chromatin conformation capture technologies present an attractive avenue for the application of graph-based methods, and so the rest of the publications covered in this thesis will focus in on the use cases of topological studies in interpreting tissue-based Hi-C (high-throughput chromatin conformation capture) data. Compared to gene regulatory networks, chromatin interaction networks are substantially larger, denser and more directly tied into the observed behavior of biomolecules, and so the layer of abstraction that network biology offers works very well here, as adjacency matrices can be more or less formed 1-to-1 out of processed Hi-C datasets.

Our work on chromatin interaction networks begins with the Javierre dataset of tissue-based Hi-C interaction networks in 17 different blood cell types (Javierre et al. 2016). This well-annotated, readily available public dataset was highly suitable for our purposes due to being easily interpreted, fully processed and usable almost directly due to its modest size (by the standards of molecular biology experiments) and convenient format (interactions are handily marked by their CHiCAGO (Cairns et al. 2016) scores in accordance with the capture Hi-C processing pipeline used by the original authors). Furthermore, since all the cell types sourced in the data are sampled from healthy human subjects and are mutually interrelated according to the hematopoietic tree (the lineage of all blood cells in the human body, starting with hematopoietic stem cells that differentiate into all blood cells covered here), we had no issue obtaining a ground truth of relatedness to compare our computational results to.

The essential thrust of our work in this publication was focused on the identification of modules (connected components in the interaction graph) that are specific to particular sub-branches of the hematopoietic lineage, in contradistinction to modules that are specific to only one cell type or modules that are common in all cell types. The purpose of this was to hopefully identify particular chromatin structures involved in cell type differentiation. Because most of the tissue interaction maps contain in excess of 150 000 interactions, both finding components and deciding their significance needed to be achieved algorithmically.

As implied previously, quite a few interactions between cell types overlap in the Javierre dataset (for a total of ~700,000 unique interactions). Therefore we constructed a large graph for each chromosome (there were relatively few interchromosomal interactions and we followed the literature at the time

(Lajoie, Dekker, and Kaplan 2015) in excluding them as likely insignificant to our analysis) where we labeled the edges and vertices depending on which cell types they constitute a sufficiently significant (that is, possessing a CHiCAGO score (Cairns et al. 2016) of 5 or higher) interaction. Then we implemented an algorithm for finding network components (FindNetworkComponents) in these graphs by using a breadth-first search of a binomial tree of all cell type combinations, a tractable enough solution for our purposes. In addition to this, we searched for components that were not too large or too small for a sensible analysis (between 10 and 100 vertices in size) and preserved more than 75% of their edges in a small subset of our cell types while failing to retain them in the broader dataset. To facilitate selection of such components for further analysis, we created the SignificanceScore which ranked all components found in our data according to these criteria (for an example of such a component see Fig. 2) and identified components of particular interest that we then could employ further analysis methods on. Our chosen algorithm performed adequately at this task, albeit it should be said that the exponential time complexity of finding network components depending on cell type count may mean it would not be efficient for much larger datasets.

Our next step in identifying if these components genuinely identified any kind of pattern of gene expression or differentiation was to check for the presence of coordinated gene expression via BLUEPRINT RNA-seq data (Adams et al. 2012) and FANTOM5 CAGE (Noguchi et al. 2017) data, which could feasibly tell us whether there was heightened gene expression within the components, but we could not manage the relatively low coverage of the dataset in comparison to our set of vertices (but would return to this thread later in subsequent publications). Instead, we utilized both the Enrichr web tool (Kuleshov et al. 2016) to test a subset of our highest scoring components as well as developed a method to analyze the variance of ChromHMM (Ernst and Kellis 2017) annotations denoting putative chromatin states based on epigenetic data to see if our selected components appeared to contain elevated amounts of activity that would be consistent with the decisive role in gene expression we intended to locate.

Results

The component structure of our Hi-C dataset proved amenable to our analysis, with some important caveats. Firstly, a substantial portion of the graph formed one large connected component containing half or more of the vertices overall, with the mid-size components we are looking for constituting a relative minority of all vertices in the dataset. This is not necessarily a flaw, as the remainder that were collected nevertheless constituted several thousand components with a somewhat even spread of sizes between 10 and 100 vertices in all chromosomes surveyed (the Y chromosome was excluded from the study

on account of having almost no interactions) and were entirely suitable for further analysis.

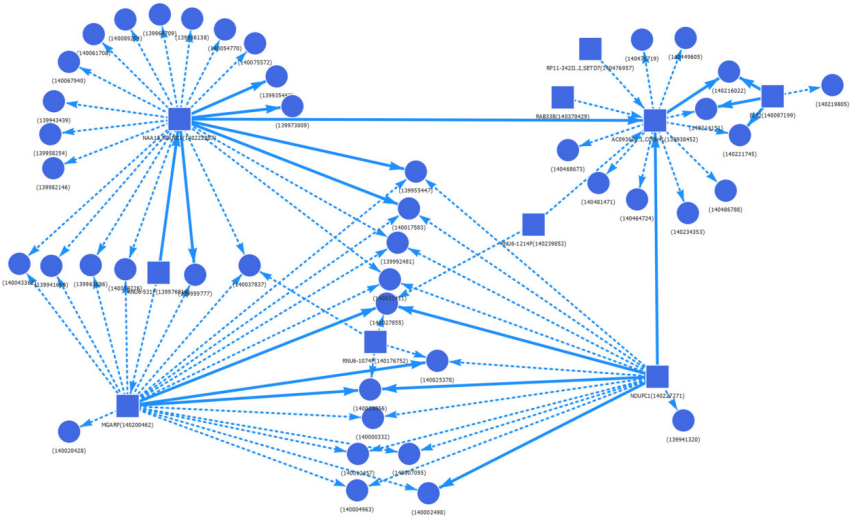

Figure 2. Example of a highly significant component under our chosen rubric. Solid lines are retained when adding tissue types outside of the main set, dotted lines are lost when doing the same.

The biological validation of our components, which involved examining both gene regulator enrichment and chromatin annotation enrichment in components compared to the dataset as a whole, overall appeared to show modest increases in chromatin annotations of all kinds, including quiescent "dead zones" and repressed areas of chromatin as well as some associations with particular transcription factors. This seemed to indicate that there was indeed activity aggregating within these connected components, but it was less than effectively refined by the particular kind of connected component we were examining. Therefore, two prime concerns emerged from this line of analysis. The first of these was that the necessity of obtaining a more specific query to sort our connected components by, such as a specific kind of subgraph or graphlet. Likely candidates for these would be highly interconnected central elements to our components, such as the starlike aggregations produced by particular locations represented by nodes with the highest node degrees, so a more detailed examination of the topology of our components would be required in the future. The second concern was that our selected components would benefit from a broader sphere of comparison to the rest of the dataset, to identify whether there is not either a broader or narrower category of component that

would be more enriched desirable properties such as gene activity. With these two concerns in mind, we moved on to more detailed topological analyses of this dataset in the following two publications.

3. PUBLICATIONS III AND IV – GRAPH-BASED CHARACTERIZATIONS OF CELL TYPES AND FUNCTIONALLY RELATED MODULES IN PROMOTER CAPTURE HI-C DATA, CHARACTERISTIC TOPOLOGICAL FEATURES OF PROMOTER CAPTURE HI-C NETWORKS

Having previously addressed the basic topology of components in our chosen dataset, we then began working on additional strategies for isolating interesting topological properties from the large graphs that our Hi-C matrices produce across their set of tissues. In the following two publications we explore the possibilities of topological metrics in differentiating tissue types from one another, primarily in the form of counting different kinds of graphlets occurring in our examined tissue types. Publications III and IV significantly overlap in subject matter, and so will be covered here in one chapter.

It is important to outline a particular property of the data studied here, which is directionality. By default, high-throughput chromatin conformation capture data cannot be said to be biologically directional because a chromatin contact is by nature reciprocal, that is, if a chromatin region *a* is said to be in close proximity with chromatin region *b*, chromatin region *b* is consequently at the same close proximity to chromatin region *a*. However, the Babraham Institute dataset we began our work with has an additional property of interest in that it is a promoter capture Hi-C dataset. Promoter capture Hi-C datasets are distinct from most Hi-C datasets in that their contact list is constrained to only include contacts involving at least one promoter region (pre-selected as part of the experiment) (Mifsud et al. 2015). These are known as "baits" while the other contact is known as the "other end" (a bait region can connect to another bait region, which is what makes larger connected components possible. Since contacts always involve at least one bait region (and are listed as such in the dataset with the bait region unambiguously marked), they can be rendered as directed edges, which opens up potential variations in topology that we examined in more detail here.

The basis of our approach here was the creation of 57 metrics that measured, among other things, the number of various kinds of connected components (connected components, bi-connected components, strongly connected components and cliques), the average or maximum edge and vertex counts in these components and sub-variations thereof (for a full explanation of the different metrics see the publications). These *Base57* metrics included most

permutations of these counts for every kind of connected component surveyed, which naturally meant that many of these would be mutually redundant in terms of predicting differences in tissue types (for example, metrics that measure the number of vertices and the number of edges for a particular kind of component would naturally be highly correlated compared to other metrics). This led to an immediate narrowing of the *Base57* set to a smaller *Base34* set of metrics by removing obvious redundancies (with correlation coefficients exceeding 0.93 in practice), which were then subjected to further refinement via stepwise regression down to a set of 11 (*Base11*) metrics in Publication III, then down to 6 (*Base6*) in Publication IV to obtain maximal predictive value with the fewest variables (as evaluated by the Akaike information criterion or AIC) in comparison with a set of bespoke distance metrics designed for our study.

Plotting out distance between cell types is in itself a challenge because there is no absolute way to gauge how different two types of cells are, which means that a certain degree of abstraction is required. In our case we made use of several distance measures to compare our metrics to, some of which came from the dataset itself (D_{cont} and D_b, both of which are based on a hierarchically clustered tree of cell types based on the contact matrices from the study) while others were computed based on the hematopoietic tree (D_4, D_A, D_B, D_C). Of these, D_A, D_B, D_C and D_b are binary measures based on whether two cell types are within the same subgroup of cell types or not, D_4 is a more complex measure wherein cell types are sorted into 5 subgroups with different distances between them and D_{cont} is directly computed from the cell type hierarchical clustering from the original publication the dataset was sourced from as depicted in Fig. 3 (Javierre et al. 2016). These six distance measures were used to optimize our set of metrics in both publications in order to obtain reliable comparisons for how they performed in discriminating between cell types.

Finally, we also analyzed the component structure further (continuing our work from Publication II) both through the previously used method of Enrichr manual transcription factor enrichment analysis (Kuleshov et al. 2016) as well as the incorporation of FANTOM5 expression profiles (Noguchi et al. 2017) into our components. FANTOM5 expression profiles were a necessary addition to our testing approach in order to discover an actual link between our selected chromatin topological patterns and gene expression events, if one genuinely exists. The coverage here remained rather low, with a relative minority of promoter regions having FANTOM5 expressions attached (covering 18% of vertices in 11 of 17 cell types for which data was available), but this was enough to make some determinations about the comparative gene activity in the cell types covered by the FANTOM5 data as far as it pertains to differences between components in different cell types. To get an idea of whether there was substantial coordination of transcription start site activity within cliques compared to the dataset as a whole, we calculated correlations between pairs of

transcription sites within particular cliques and compared them statistically to sites outside of those cliques, also considering transcriptional activities observed in closely related vs. unrelated cell types.

4.1. Results

Our chosen approach was moderately successful in unpicking more of the complexities in our chosen dataset, as the metrics used, tested and refined showed some ability in distinguishing particular cell types from one another, although this was not entirely consistent between chromosomes. Chromosome-specific patterns of correlation in particular metrics confounded the analysis, and while there were particular patterns observed such as the heightened significance of bi-connected components, our topological approach did not necessarily provide any explanation as to why these specific patterns would be significant, which made validation doubly important as a subsequent step. This was complicated further by our chosen distance metrics – while it was possible to get fairly exact distances with D_{cont}, this distance measure in particular was derived from the dataset itself and, while it was substantially different enough from what our metrics were measuring, it nevertheless had unavoidable bias compared to the hematopoietic tree metrics, which were less precise and harder to evaluate in concrete terms. This created obvious issues that we would attempt to solve in later work and ultimately led to us abandoning the metric approach for later work, including Publication V.

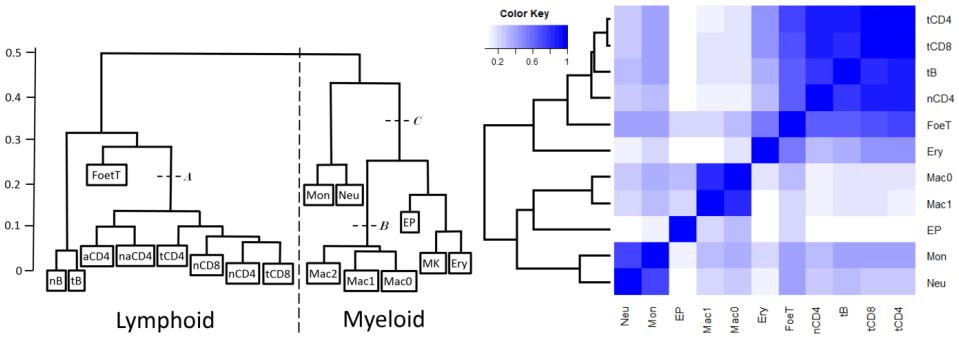

Figure 3. The D_{cont} measure of relatedness visualized as a heatmap, obtained by measuring distances across a clustered tree of cell types (Javierre et al. 2016).

Another matter to consider is that our methodology, while theoretically suitable for a wide variety of Hi-C datasets, had thus far been tested primarily

just on the single dataset from the Babraham Institute about the lineage of blood cells. One reason for this was that the blood cell pcHi-C dataset was comprehensive and made up of healthy human cells with a clearly established relation between them, which simplified biological interpretation, unlike many other Hi-C datasets that focus on cell lines made up of either cancer cells or similar immortal cell cultures that may have additional peculiarities in chromatin structure that need accounting for. Additionally, because our method was developed for pcHi-C data, the much larger size and density of other Hi-C datasets, including gold-standard high-resolution datasets used widely in the field (Rao et al. 2014), presented difficulties in efficiently computing connected components for them in a practical amount of time.

Nevertheless, the information gathered in these publications demonstrated the usefulness of focusing in on narrower topological definitions such as cliques, which we would continue to explore as potentially likely candidates for significant chromosomal structures. Of particular significance is the inclusion of expression profiles here, which for the first time provide a substantial enough basis by which to judge gene expression in our components in a substantial way. While the approach used needed additional refinement to be fully suitable to the data, it nevertheless provided reasonable evidence of actual gene regulatory events occurring within our chosen components. These methodological improvements would form the basis of our future work.

4. PUBLICATION V – TRANSCRIPTIONAL HUBS WITHIN CLIQUES IN ENSEMBLE HI-C CHROMATIN INTERACTION NETWORKS

With a more elaborate understanding of how to characterize both connected components and smaller topological patterns (much of the intervening work on tissue similarity and metrics remains unpublished) we would continue to refine and streamline our analytical approach, at last achieving some success with a clique-specific analysis of chromatin conformation data covered in Publication V, which is itself an elaboration on an earlier conference paper (Melkus et al. 2023).

Focusing on cliques as the primary object of interest has a number of benefits both methodologically and computationally over the previous metric-based approach that used connected components more generally. Firstly, methodological concerns about the soundness of how our connected components were selected, calculated and further analyzed were no longer in play. Secondly, our selection of the narrower topological category of cliques meant that the function of connected components in narrowing our search down to computationally tractable boundaries was fulfilled more efficiently by the category of cliques of size 3 (C3), a triangle of vertices fully connected by edges, which are a simple clique that nevertheless fully encompasses all larger aggregations of chromatin in the data. Thirdly, cliques are much more specific and rarer in the datasets studied, therefore making it easier to identify particular modules of regulation by virtue of cliques representing local maximums of connectivity (or, biologically speaking, functionally relevant aggregations of active loci) within a given cluster of chromatin.

Furthermore, we successfully applied our approach to new datasets principally obtained from the 3DIV database which included a number of Hi-C datasets for human tissues, notably a reasonably exhaustive set of both promoter capture Hi-C (for 27 different tissues) (Jung et al. 2019) and non-capture Hi-C experiments (for 20 tissues matching our pcHi-C set) (Kim et al. 2021) that we made use of for an efficient comparison of our methods across different types of data. Unlike our previous data, these particular tissue samples did not have an immediate relatedness comparable to that of the blood cell pcHi-C dataset, and while we made some effort to establish tissue similarities based on expression profiles (Manatakis, VanDevender, and Manolakos 2021), the data ultimately proved incompatible with a metric-based approach.

Our approach also was changed in terms of graph generation. Whereas for our blood cell pcHi-C data we created graphs of directed edges based on

the bait-other end dichotomy in the data, the non-capture Hi-C dataset did not have any reason to make such a distinction, and to make sure the results would be easier to compare, we rendered both datasets as undirected graphs. For each chromosome in each tissue we constructed graphs G_t, where vertices V_t are genomic regions found in the Hi-C contact matrix and edges E_t are contacts that met our filtering criteria. Since these datasets were not processed according to the CHiCAGO pipeline, we filtered contacts by their assigned p-values instead, using cutoff points chosen to optimize the size of our overall graph, which in practice meant a much higher cutoff point for the denser Hi-C data than for the pcHi-C data. For each data type the resulting graphs were then assembled into a comprehensive graph where each edge was additionally labeled with the tissues that it has been found in. We proceeded to find cliques in these graphs, noting their positioning on the chromosome as well as the edges and vertices involved.

To validate the functional significance of the cliques we located we employed a wide range of data. Firstly, by adding Ensembl genes to our contact map at their appropriate positions in the genome we were able to perform gene ontology enrichment analysis on the contacts within cliques as opposed to contacts in the dataset as a whole using GOEA (Klopfenstein et al. 2018). Secondly, we obtained a wide range of chromatin state modeling data (derived from epigenetic assays and assigned via hidden Markov model to the genome, denoting features such as transcription start sites, enhancers and more) from Roadmap Epigenomics (Abascal et al. 2020) covering much of our tissue data and were able to use this to analyze enrichment in functional annotations in cliques vs. the genome as a whole. Thirdly, we assigned expression profiles from the Genotype-Tissue Expression atlas (GTEx) (Lonsdale et al. 2013) and FANTOM5 (Noguchi et al. 2017) to our vertices, with the additional refinement that we did not require this positional expression to ideally match up with the vertex, which is consistent with findings in the literature that gene expression and chromatin conformation do not need to match perfectly in coordinates to have a functional link (Wurmser and Basu 2022). Finally, to help us narrow down the results, we also brought in additional annotations from the Encyclopedia of DNA Elements (ENCODE) project (Jou et al. 2019), most notably binding data for RNA polymerase II, to help further refine our analysis of gene expression in cliques by narrowing it down to specifically active loci. This set of approaches allowed us to obtain considerably more reliable and comprehensive results than previously.

5.1. Results

The essential finding with regards to cliques of size 3 in our Hi-C data was the noticeable prevalence of high concentrations of cliques in particular

chromosomal locations, which was consistent with observed "chromatin hotspots" from the literature indicative of active gene clusters (Liu et al. 2019). Isolating cliques in particular proved an effective way to analyze the densest parts of chromatin clusters and identify fairly robust patterns of expression and epigenetic markings within them.

Notably, our multi-pronged validation strategy showed promising results along a number of lines. Firstly, our gene ontology enrichment analyses demonstrated the presence of several previously known gene clusters (such as the HLA cluster or olfactory receptor protein genes) and several less obvious aggregations that may very well fulfill a similar function. Secondly, we saw noticeable enrichment of certain chromatin annotations within our cliques, specifically those of transcription start sites and active transcription while curiously seeing a much lesser proportion of enhancer annotations (see Fig. 4), suggesting that our identified gene modules match up more closely to the concept of transcription factories rather than promoter-enhancer interactions. Thirdly, our analysis of gene expression produced a surprisingly solid result in the form of a clear difference in gene expression in both FANTOM5 and GTEx datasets when considering cliques as opposed to contacts in the dataset as a whole, an effect that became even more apparent when analyzing only RNA polymerase II binding sites where active transcription is reasonably plausibly occurring in the tissues in question. All of these lines of evidence pointed toward our cliques somewhat effectively filtering out clusters of coordinated transcription rather than other kinds of cis-regulatory elements.

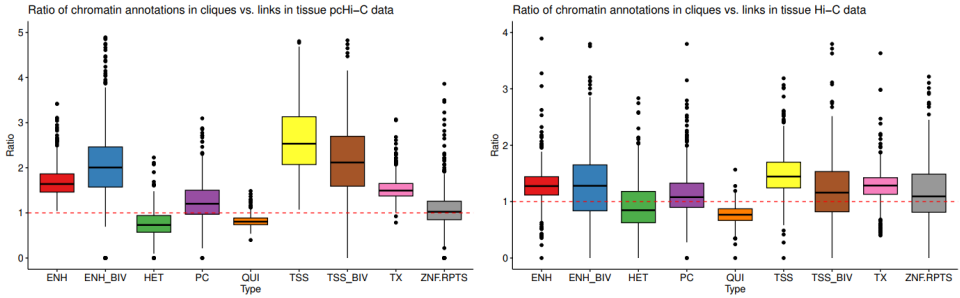

Figure 4. Enrichment of chromatin annotations in cliques vs. links in our tissue data, measured as a ratio in all chromosomes. The red dotted line at ratio 1 indicates the border line between enrichment and depletion.

All of this opened further avenues of analysis in terms of characterizing these cliques more precisely, as even with the narrower selection criteria of cliques the overall clusters found were still challenging to pick out specific

elements from. Fortunately, the system of annotations applied here fit easily with positional formatting standards in the field of computational biology, which meant that further information on a variety of genomic features will be easy enough to come by in the future. However, to more conclusively establish patterns of expression in cells such as these, we will eventually need to move on to more granular kinds of chromatin conformation data – most notably single cell Hi-C datasets which, unlike ensemble (or bulk) Hi-C data, afford a much more immediate, less averaged view of chromatin architecture. However, our current use case presents a successful implementation of a graph-based topological approach in analyzing Hi-C data, finding and successfully identifying chromatin hubs within previously known biological categories based on graph topology alone. Further innovation along these lines is definitely possible, and topological analysis of newer single-cell datasets is likely to bear further novel results.

CONCLUSION

The essential implementations of graph-based models for biomolecular data featured in this study were successful in their primary aim, which is to facilitate a systematic understanding of how to interpret interactions between biomolecules. The key findings included in this thesis can be summarized as follows:

- Our topological approach was successful in identifying symmetry in paralogous gene pairs via the use of bi-fan units, showing that overlap in motifs was indeed significantly elevated in duplicated genes compared to either random genes or out-of-species paralogs. This agrees with the theoretical assertion that duplicated genes retain most of their interactions, which are then slowly lost over time as the genes diverge.
- Connected components in chromatin interaction networks provide numerous avenues for topological investigation, including but not limited to possibilities of studying cell differentiation along topological lines, identifying functional modules in chromatin and locating hotspots of regulatory activity.
- Larger topological units of study (such as sizable connected components in chromatin interaction networks) suffer from being overly large and non-specific in terms of our ability to categorically identify their true composition and biological significance as well as challenging to replicate in other data sets, so deriving more useful conclusions about the data necessarily requires a more focused investigation of particular topological elements such as cliques.
- Genomic datasets such as Hi-C benefit considerably from positional genomic information such as epigenetic markers, positional gene expression data, the locations of cis-regulatory elements as well as genes, all of which can be somewhat neatly integrated into a graph-based model and subjected to further study.
- Cliques, a topological element that denotes a strong aggregation of chromatin, are the most promising simple topological element for further study, given that they are not only somewhat simple to find and assess in both directed and undirected graphs, but also align very well with existing molecular biology concepts such as transcription factories and demonstrate a very noticeable increase in transcription, an enrichment in relevant epigenetic marks as well as related measures such as RNA polymerase II binding in several unrelated datasets.

The findings of this thesis can be built upon in the future, not just with further topological refinements and usage of newer data (such as single-cell high throughput chromatin conformation capture datasets) but also with more sophisticated means of integrating biological information. Overall, a substantial body of work is available in the field, particularly with the development of more and more new methods and the availability of more reliable data, and the conclusions described above can be carried forward into many avenues of future research.

REFERENCES

Abascal, Federico, Reyes Acosta, Nicholas J. Addleman, Jessika Adrian, Veena Afzal, Rizi Ai, Bronwen Aken, et al. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." *Nature* 583 (7818): 699–710. https://doi.org/10.1038/s41586-020-2493-4.

Adams, David, Lucia Altucci, Stylianos E Antonarakis, Juan Ballesteros, Stephan Beck, Adrian Bird, Christoph Bock, et al. 2012. "BLUEPRINT to Decode the Epigenetic Signature Written in Blood." *Nature Biotechnology* 30 (3): 224–26. https://doi.org/10.1038/nbt.2153.

Alon, Uri. 2007. "Network Motifs: Theory and Experimental Approaches." *Nature Reviews Genetics* 8 (6): 450–61. https://doi.org/10.1038/nrg2102.

Babu, M. Madan, Nicholas M. Luscombe, L. Aravind, Mark Gerstein, and Sarah A. Teichmann. 2004. "Structure and Evolution of Transcriptional Regulatory Networks." *Current Opinion in Structural Biology* 14 (3): 283–91. https://doi.org/10.1016/j.sbi.2004.05.004.

Barabási, Albert László, and Zoltán N. Oltvai. 2004. "Network Biology: Understanding the Cell's Functional Organization." *Nature Reviews Genetics* 5 (2): 101–13. https://doi.org/10.1038/nrg1272.

Broido, Anna D., and Aaron Clauset. 2019. "Scale-Free Networks Are Rare." *Nature Communications* 10 (1): 1017. https://doi.org/10.1038/s41467-019-08746-5.

Byrne, Kevin P., and Kenneth H. Wolfe. 2005. "The Yeast Gene Order Browser: Combining Curated Homology and Syntenic Context Reveals Gene Fate in Polyploid Species." *Genome Research* 15 (10): 1456–61. https://doi.org/10.1101/gr.3672305.

Cairns, Jonathan, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, et al. 2016. "CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C Data." *Genome Biology* 17 (1): 127. https://doi.org/10.1186/s13059-016-0992-2.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. "Capturing Chromosome Conformation." *Science* 295 (5558): 1306–11. https://doi.org/10.1126/science.1067799.

Ernst, Jason, and Manolis Kellis. 2017. "Chromatin-State Discovery and Genome Annotation with ChromHMM." *Nature Protocols* 12 (12): 2478–92. https://doi.org/10.1038/nprot.2017.124.

Harrison, Peter W, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Arne Becker, et al. 2024. "Ensembl 2024." *Nucleic Acids Research* 52 (D1): D891–99. https://doi.org/10.1093/nar/gkad1049.

Javierre, Biola M., Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, Paula Freire-Pritchett, et al. 2016. "Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters." *Cell* 167 (5): 1369-1384.e19. https://doi.org/10.1016/j.cell.2016.09.037.

Jou, Jennifer, Idan Gabdank, Yunhai Luo, Khine Lin, Paul Sud, Zachary Myers, Jason A. Hilton, et al. 2019. "The ENCODE Portal as an Epigenomics Resource." *Current Protocols in Bioinformatics* 68 (1). https://doi.org/10.1002/cpbi.89.

Jung, Inkyung, Anthony Schmitt, Yarui Diao, Andrew J. Lee, Tristin Liu, Dongchan Yang, Catherine Tan, et al. 2019. "A Compendium of Promoter-Centered Long-Range Chromatin Interactions in the Human Genome." *Nature Genetics* 51 (10): 1442–49. https://doi.org/10.1038/s41588-019-0494-8.

Kellis, Manolis, Bruce W. Birren, and Eric S. Lander. 2004. "Proof and Evolutionary Analysis of Ancient Genome Duplication in the Yeast Saccharomyces Cerevisiae." *Nature* 428 (6983): 617–24. https://doi.org/10.1038/nature02424.

Kim, Kyukwang, Insu Jang, Mooyoung Kim, Jinhyuk Choi, Min-Seo Kim, Byungwook Lee, and Inkyung Jung. 2021. "3DIV Update for 2021: A Comprehensive Resource of 3D Genome and 3D Cancer Genome." *Nucleic Acids Research* 49 (D1): D38–46. https://doi.org/10.1093/nar/gkaa1078.

Klopfenstein, D. V., Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, et al. 2018. "GOATOOLS: A Python Library for Gene Ontology Analyses." *Scientific Reports* 8 (1): 10872. https://doi.org/10.1038/s41598-018-28948-z.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. "Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update." *Nucleic Acids Research* 44 (W1): W90–97. https://doi.org/10.1093/nar/gkw377.

Lajoie, Bryan R, Job Dekker, and Noam Kaplan. 2015. "The Hitchhiker's Guide to Hi-C Analysis: Practical Guidelines." *Methods* 72 (January): 65–75. https://doi.org/10.1016/j.ymeth.2014.10.031.

Larsen, Simon J., Richard Röttger, Harald H.H.W. Schmidt, and Jan Baumbach. 2019. "E. Coli Gene Regulatory Networks Are Inconsistent with Gene Expression Data." *Nucleic Acids Research* 47 (1): 85–92. https://doi.org/10.1093/nar/gky1176.

Lieberman-Aiden, Erez, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Lima-Mendez, Gipsi, and Jacques van Helden. 2009. "The Powerful Law of the Power Law and Other Myths in Network Biology." *Molecular BioSystems* 5 (12): 1482. https://doi.org/10.1039/b908681a.

Liu, Li, Qian-Zhong Li, Wen Jin, Hao Lv, and Hao Lin. 2019. "Revealing Gene Function and Transcription Relationship by Reconstructing Gene-Level Chromatin Interaction." *Computational and Structural Biotechnology Journal* 17: 195–205. https://doi.org/10.1016/j.csbj.2019.01.011.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nature Genetics* 45 (6): 580–85. https://doi.org/10.1038/ng.2653.

Manatakis, Dimitris V, Aaron VanDevender, and Elias S Manolakos. 2021. "An Information-Theoretic Approach for Measuring the Distance of Organ Tissue Samples Using Their Transcriptomic Signatures." Edited by Pier Luigi Martelli. *Bioinformatics* 36 (21): 5194–5204. https://doi.org/10.1093/bioinformatics/btaa654.

Melkus, Gatis, Sandra Silina, Andrejs Sizovs, Peteris Rucevskis, Lelde Lace, Edgars Celms, and Juris Viksna. 2023. "Clique-Based Topological Characterization of Chromatin Interaction Hubs." In , 476–86. https://doi.org/10.1007/978-981-99-7074-2_38.

Mifsud, Borbala, Filipe Tavares-Cadete, Alice N. Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W. Wingett, et al. 2015. "Mapping Long-Range Promoter Contacts in Human Cells with High-Resolution Capture Hi-C." *Nature Genetics* 47 (6): 598–606. https://doi.org/10.1038/ng.3286.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. "Network Motifs: Simple Building Blocks of Complex Networks." *Science* 298 (5594): 824–27. https://doi.org/10.1126/science.298.5594.824.

Noguchi, Shuhei, Takahiro Arakawa, Shiro Fukuda, Masaaki Furuno, Akira Hasegawa, Fumi Hori, Sachi Ishikawa-Kato, et al. 2017. "FANTOM5 CAGE Profiles of Human and Mouse Samples." *Scientific Data* 4 (1): 170112. https://doi.org/10.1038/sdata.2017.112.

Pancaldi, Vera. 2021. "Chromatin Network Analyses: Towards Structure-Function Relationships in Epigenomics." *Frontiers in Bioinformatics* 1 (October). https://doi.org/10.3389/fbinf.2021.742216.

———. 2023. "Network Models of Chromatin Structure." *Current Opinion in Genetics & Development* 80 (June): 102051. https://doi.org/10.1016/j.gde.2023.102051.

Pržulj, Nataša. 2007. "Biological Network Comparison Using Graphlet Degree Distribution." *Bioinformatics* 23 (2): e177–83. https://doi.org/10.1093/bioinformatics/btl301.

Rao, Suhas S.P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7): 1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

Reece-Hoyes, John S., Carles Pons, Alos Diallo, Akihiro Mori, Shaleen Shrestha, Sreenath Kadreppa, Justin Nelson, et al. 2013. "Extensive Rewiring and Complex Evolutionary Dynamics in a C.Elegans Multiparameter Transcription Factor Network." *Molecular Cell* 51 (1): 116–27. https://doi.org/10.1016/j.molcel.2013.05.018.

Sarajlić, Anida, Noël Malod-Dognin, Ömer Nebil Yaveroğlu, and Nataša Pržulj. 2016. "Graphlet-Based Characterization of Directed Networks." *Scientific Reports* 6 (1): 35098. https://doi.org/10.1038/srep35098.

Sizovs, Andrejs, Gatis Melkus, Peteris Rucevskis, Sandra Silina, Lelde Lace, Edgars Celms, and Juris Viksna. 2024. "A Technique for Preserving Network Structure in Randomized Hi-C Data." *Journal of Bioinformatics and Computational Biology* 22 (05). https://doi.org/10.1142/S0219720024400018.

Sizovs, Andrejs, Sandra Silina, Gatis Melkus, Peteris Rucevskis, Lelde Lace, Edgars Celms, and Juris Viksna. 2024. "Exploration and Visualization Methods for Chromatin Interaction Data." In , edited by Wei Peng, Zhipeng Cai, and Pavel Skums, 101–13. Singapore: Springer Nature Singapore.

Sorrells, Trevor R., and Alexander D. Johnson. 2015. "Making Sense of Transcription Networks." *Cell* 161 (4): 714–23. https://doi.org/10.1016/j.cell.2015.04.014.

Stone, Lewi, Daniel Simberloff, and Yael Artzy-Randrup. 2019. "Network Motifs and Their Origins." Edited by Ruth Nussinov. *PLOS Computational Biology* 15 (4): e1006749. https://doi.org/10.1371/journal.pcbi.1006749.

Teixeira, Miguel Cacho, Romeu Viana, Margarida Palma, Jorge Oliveira, Mónica Galocha, Marta Neves Mota, Diogo Couceiro, et al. 2023. "YEASTRACT+: A Portal for the Exploitation of Global Transcription Regulation and Metabolic Model Data in Yeast Biotechnology and Pathogenesis." *Nucleic Acids Research* 51 (D1): D785–91. https://doi.org/10.1093/nar/gkac1041.

Tierrafría, Víctor H., Claire Rioualen, Heladia Salgado, Paloma Lara, Socorro Gama-Castro, Patrick Lally, Laura Gómez-Romero, et al. 2022. "RegulonDB 11.0: Comprehensive High-Throughput Datasets on Transcriptional Regulation in Escherichia Coli K-12." *Microbial Genomics* 8 (5). https://doi.org/10.1099/mgen.0.000833.

Vandereyken, Katy, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. 2023. "Methods and Applications for Single-Cell and Spatial Multi-Omics." *Nature Reviews Genetics* 24 (8): 494–515. https://doi.org/10.1038/s41576-023-00580-2.

Ward, Jonathan J., and Janet M. Thornton. 2007. "Evolutionary Models for Formation of Network Motifs and Modularity in the Saccharomyces Transcription Factor Network." *PLoS Computational Biology* 3 (10): 1993–2002. https://doi.org/10.1371/journal.pcbi.0030198.

Wurmser, Annabelle, and Srinjan Basu. 2022. "Enhancer-Promoter Communication: It's Not Just About Contact." *Frontiers in Molecular Biosciences* 9 (April). https://doi.org/10.3389/fmolb.2022.867303.