Phylogenetic Inference from Enumeration of Automorphism Orbits of Graphlets

Galen Michael Seilis
[redacted]

December 7, 2020

Contents

1 Introduction

This section provides some background information about the history and concepts in the project, leaving the project details for later sections. It covers similar material to the proposal, including the concept of phylogenetic inference, interactomics, and graph theory.

1.1 Evolution and Phylogenetic Inference

People have speculated about the diversity of life on Earth since antiquity, but it wasn't until the mid-1700's that Charles Darwin made significant progress with his *magnum opus "On the Origin of Species"* in which he layed out his principle of natural selection.[1] While a socially-controversial work for its time, it became the dominant framework for scientists to generate hypotheses about biological systems.[1]

The notion of ancestry that is evident from the facts of reproduction suggest a hierarchy of similarity among organisms, combined with observations that siblings and cousins resemble each other more than strangers. This suggested to naturalists even before Charles Darwin that there was an underlying hierarchy of relatedness, although it was without contention from biblical creationists. Phylogenetics is the study of the history and relatedness of life on Earth caused by mutational, reproductive, random, and selective processes. A phylogram, phylogenetic tree, or evolutionary tree, is a hierarchical representation of the relatedness among organisms or clades.[1] These trees are tested as hypotheses through hypothetico-deductive methods[2], with the agreement of the trees to data being evaluated through maximal parsimony, maximum likelihood, or Bayesian methods.[3] There is a hypothetically a 'universal tree' showing the relatedness of all organisms, termed the *tree of life*, however it is now understood that this would only be a useful approximation. Why the relatedness of life on Earth is not entireyl hierarchical is due to various forms of horizontal gene transfer. Some of these horizontal transfers of genetic material include conjugation, transduction, transformation, and endosynbiosis.[4, 5, 6, 7] However, it should be emphasized that for much of the tree of life that includes sexual organisms, the vast majority of genetic variation is explained by vertical transfer.

1.2 Interactomics

Interactomics is the study of all interactions within biological systems. These *interactions* can include direct physical interactions such as hydrogen bonding between proteins, as well as abstract interactions like gene regulation that are biologically implemented by the cell. An *interactome* is the set of all such interactions within a cell.[8] Some interactions are symmetric, like binding between molecules, where both entities participating in the interaction do so in an equivalent way. Others can be asymmetric, like the regulation of genes where often one gene controls the other but not the other way around. Many genes regulate each other in negative feedback loops where the production of one inhibits the production of the other. The Biological General Repository for Interaction Datasets (BioGrid) is a public access database containing nearly 2 million protein-protein interactions, genetic interactions, chemical associations, and post-translation modifications for 71 model organisms and viruses.[9] Figure 1 shows the available organisms as a tree representation.[2]

[1] Note that evolutionary trees are representing an abstract representation of lineages of populations or metapopulations. They are not equivalent to a pedigree, showing individual-level histories of reproduction. This point is easily confused when the leafs of an evolutionary tree are individuals.

[2] Figure 1 is rendered as a PDF, so it should be possible to zoom in to inspect the diagram closely, if desired.

Interactome data offers new opportunities for statistical and network analysis of biological systems.

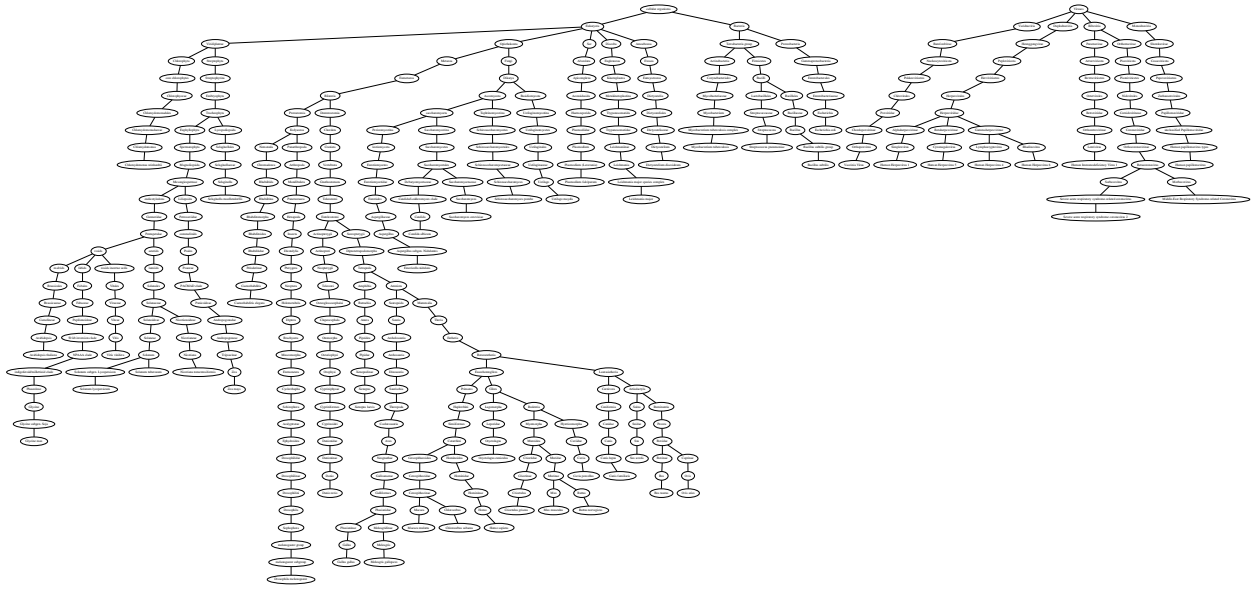


Figure 1: Tree representation of the available ogranisms in the BioGrid database. The left connected component represents cellular organisms such as plants and animals, and the right connected component includes acellular entities such as viruses. Each leaf is a species or strain, with the internal nodes representing more abstract clades from leaf-to-roof.

1.3 Graph Theory and Orbit Automorphisms

Graph theory is a branch of mathematics that has enervated both the pure and applied study of binary relations and functions ever since Leonard Euler developed its foundation in his famous analysis of the *The Seven Bridges of Konigsberg.*[10] A binary relation itself is a subset of the Cartesian product of two sets. A graph g is a tuple (V, E) where V is a set of vertices and E is a set of edges where $E \subseteq V \times V$. Thus every edge set represents a relation over the vertex set of a graph. Real-world datasets have variables with relationships that can be operationally represented as relations, which has led to discipline of network analysis or network science that combines graph theory with other methods and perspectives.

An isomorphism is a bijection that preserves structure, however what consistutes structure depends on context. Often, structure is a relation on the elements of the domain and image of such a function.[3] In the case of graph isomorphism, this is conventionally defined to preserve the property of adjacency of vertices. That is to say, graph isomorphism is a bijection between the vertices of one graph and the vertices of another graph such that if any two vertices in the domain graph are adjacent, then they will also be adjacent in the image graph. A graph automorphism is a graph isomorphism between a graph and itself, and one can define equivalence classes nodes under automorphism that are conventionally called automorphism orbits. A more detailed decomposition of the concepts is given in Figure 2.

[3]This is not *necessarily* true in more general discussions of structure-preserving maps, but category theory is beyond the scope of this report.

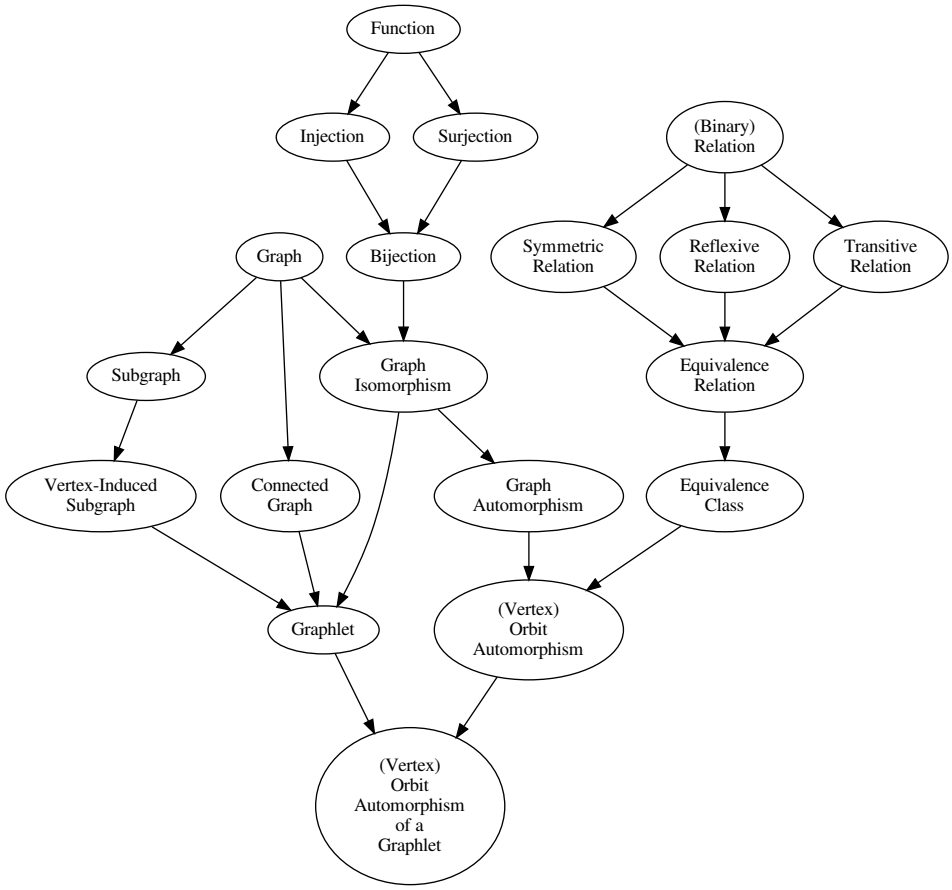


Figure 2: Mindmap of the concepts required for understanding automorphism orbits of graphlets. The most ancestral or fundamental concepts in this diagram are the notion of functions, binary relations, and graphs. These fundamental concepts are also related to each other, but for simplicity there are no paths between them. Note that this diagram does not follow any formal relations, and is rather intended as a visual guide. Errata: For two of the nodes the wording *Orbit Automorphism* should be replaced with *Automorphism Orbit*.

The concept of a graphlet has become popular as a way of describing local structure of a graph. A graphlet is a connected and vertex-induced subgraph of a graph that is a member of an equivalence class of such graphs under non-isomorphism.[11] Since a graphlet is itself a graph, they have automorphism orbits. For a given graph, an $m \times n$ matrix called *graphlet orbit matrix* can be calculated where the m rows represent the node indices, the n columns represent the automorphism orbit indices, and the entries are occupied by the count of how many times each node participates in a given automorphism orbit. This feature extraction gives a quantitative representation of graph structure that can be used in the analysis of data.[12]

1.4 Project Idea

The inspiration for this project came from a couple of studies that showed that the topologies of metabolic pathways had phylogenetic information, but they did not use graphlet enumeration.[13,

14] While metabolic networks are not the same as interaction networks *per se*, they are related through their chemical and physical meaning. For example, DNA polymerase II is an enzyme in prokaryotes responsible for replicating DNA, and requires physical association with magnesium to form its active site.[15] Graphlet counting[4] has been performed on the BioGrid database, however there was no aim to extract phylogenetic information in that study.[16] This apparent gap in the literature led to the idea of using automorphism orbits of graphlets as a feature that could be exploited for prediction of evolutionary relationships.

2 Objective

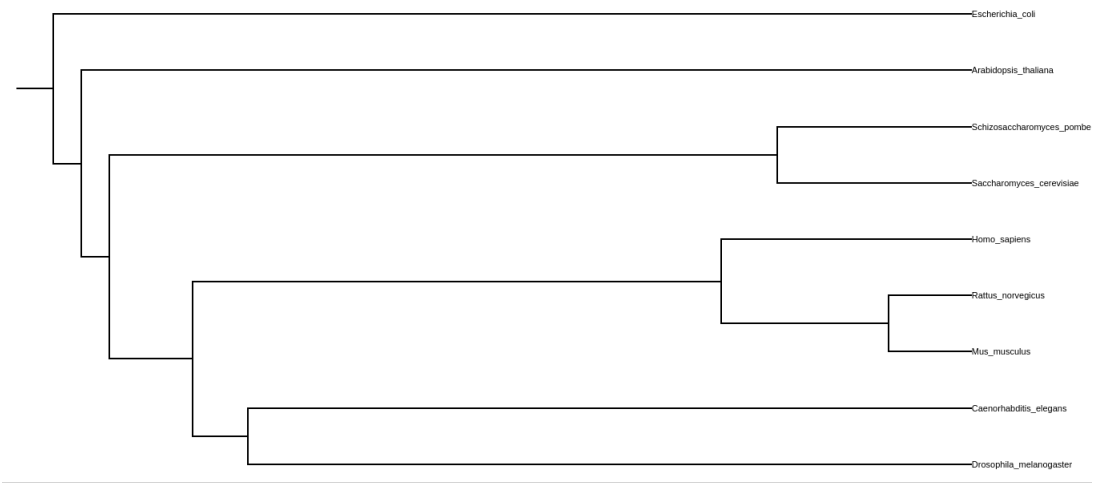
The abstract objective of this project was to test a proof on concept that an interaction network contained information useful to phylogenetic inference. The concrete objective of this project was to apply an hierarchical clustering method to graphlet orbit matrices calculated from interaction data to produce a tree that could be evaluated against a reference taxonomy for correctness.

3 Results and Discussion

This section summarizes the results of the analyses performed in the project, and then interpret what the results mean.

The agglomerative hierarchical clustering yielded an empirical tree as shown in Figure 3, which can be visually compared to the reference tree also shown in Figure 3. Visually, it is evident that many of the relationships in the reference tree are not preserved in the empirical tree. For example, the empirical tree suggests that *Homo sapiens* (humans) are more closely related to *Escherichia coli* (a common bacterial species) than they are to *Rattus norvegicus* (the Norwegian Rat). Numerous such mismatches can be made, and therefore it would seem apt to quantify how often these two trees agree or disagree.

[4] But not enumeration of automorphism orbits, insofar as I have read the literature.



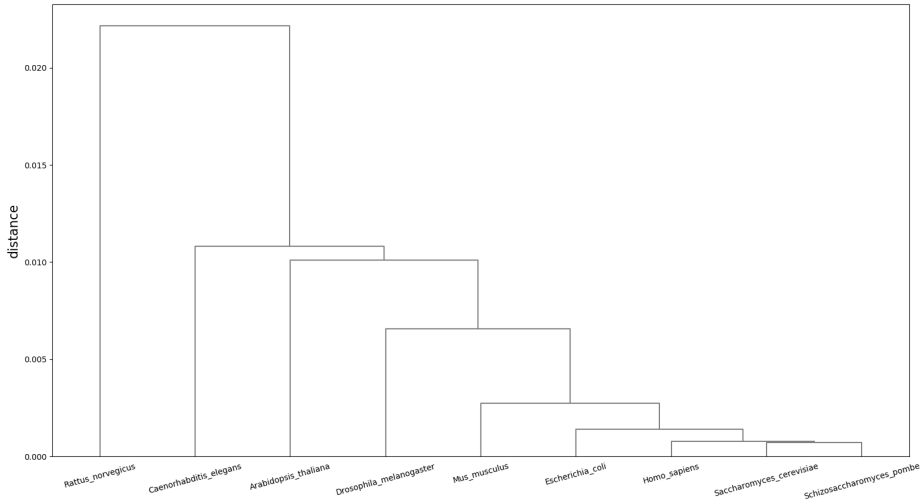


Figure 3: (top) Reference taxonomy for the species under which the analysis was performed. (bottom) Empirical dendrogram of the clustered graphlet orbit matrices under the cosine semimetric.

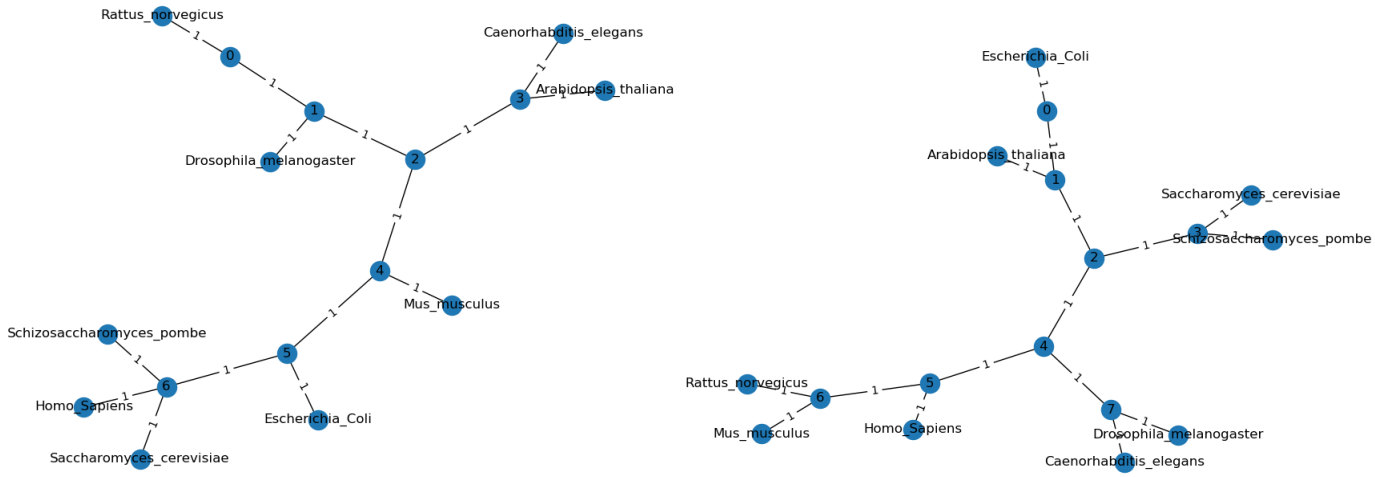


Figure 4: (left) Unrooted empirical tree with path lengths all equal to unity. (right) Unrooted reference tree with path lengths all equal to unity.

In order to quantify the agreement between the trees in Figure 3, they were unrooted and given path weights as shown in Figure 4. It remains visually evident in Figure 4 that there are incorrect relations. For any three clades, there should be two members that are more closely related to each other than they are to the third outgroup. Given the 9 clades selected for analysis, it can be enumerated how many times a such triples are in agreement with both trees, one tree and not the other, or neither. These results, along with other confusion matrix statistics, are included in Table 1 with the reference taxonomy treated as the ground truth.

Table 1: Confusion Matrix Summary from ((A,B)C) Analysis between the Estimated Taxonomy and Reference Taxonomy under the Cosine Semimetric.

Statistic	Score
total population	252.0
positive	98.0
negative	154.0
predicted condition positive	102.0
predicted condition negative	150.0
true positive	39.0
false positive	63.0
false negative	59.0
true negative	91.0
prevalence	0.39
accuracy	0.52
precision	0.38
false discovery rate	0.62
false omission rate	0.39
negative predictive value	0.61
recall	0.4
false negative rate	0.6
false positive rate	0.41
specificity	0.59
positive likelihood ratio	0.97
negative likelihood ratio	1.02
diagnotistic odds ratio	0.95
f1 score	0.39
prevalence threshold	0.5
threat score	0.24
balanced accuracy	0.49
matthews correlation coefficient	-0.01
fowlkes mallows index	0.39
bookmaker informedness	-0.01
markedness	-0.01
fishers exact test p	0.1

While there are a lot of statistics in Table 1 to be interpreted, there are a handful I would like to emphasize here. The first statistics is the accuracy, which represents the frequency of how often the empirical taxonomy agreed with the reference taxonomy on whether a given ((A,B),C) present. With an accuracy of only 52%, it is clear that the empirical model's performance is not substantially better than a coin flip. The low precision agrees with a lack of agreement between the trees. The false discovery rate was actually higher than the accuracy, indicating that the empirical tree was more likely to have a false positive than an accurate relatedness among any three taxa. The low recall indicates that the probability of the empirical tree having a true relation given that the relation exists in the reference tree is quite low. The f1-score, which is the harmonic mean of the precision and recall, was also low, giving an assurance that the model's overall performance was poor compared to chance. While not a traditional confusion matrix statistic, the Matthews correlation coefficient (given in Equation 1) also indicates that there is no association between the

reference tree and empirical tree after considering true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).[17] Lastly, a Fisher's exact test gave $\rho = 0.1$ which is not significant at $\alpha = 0.05$, implying that the values of the true positives, false positives, false negatives, and true negatives do not appear to be extremely small or larger.[18] This further indicates that the association between trees is weak.

$$\text{MCC} = \frac{(\text{TP})(\text{TN}) - (\text{FP})(\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{1}$$

While previous work indicates that metabolic networks can contain phylogenetic information[13, 14], the straightforward approach taken here on interaction networks did not seem to be related to previously-established taxonomies. This may be due to there actually being no relevant information in interaction networks that indicate ancestry, however there are confounders that do not allow for absolute confidence in this interpretation.

One reason is that the graphlet orbit matrix is a lossy representation of the original graph structures, and that potentially including additional graphlets would improve model performance. This possibility becomes increasingly difficult to test because the number of orbits grows combinatorially as a function of the size of graphlets being considered. Yet another angle to consider is whether graphlet orbit matrices serve as the best feature representation of networks for phylogenetic inference. For example, perhaps a graph embedding of fixed vector length could perform either better or worse than graphlets.

Another possibility goes right back to the data. If there are biases and errors in the data collected within the database, that could lead to potential problems later on in the analysis. Another data-oriented confounder is the usage of undirected edges rather than directed edges. While using directed edges will rapidly increase the number of orbits to calculate, is also potentially means more variation that might distinguish between networks. Perhaps seggregating the interaction networks into different subnetworks before orbit enumeration would have provided useful information.

A fourth direction to consider are the choices of distance functions and clustering algorithms. While I took an agglomerative approach here, it is possible to obtain different results with divisive approaches.[19] Perhaps a different distance function might have improved performance.[5]

Yet another consideration is the type of model and parametrization. Since modern phylogenies are constructed using Bayesian methods like resampling of trees from a prior distribution.[3]

4 Requirements

This section will concisely list the skills required to perform this project.

- General knowledge of evolutionary theory

[5] Although, I did try other distance functions such as the Euclidean distance on the lower diagonal of the orbit correlation matrix, the Euclidean distance on the orbit-aggregated-by-summation vector, and the correlation distance on the orbit-aggregated-by-summation vector. None of these did better than the cosine semimetric in terms of the statistics described above.

- Fundamental knowledge of automorphism orbits of graphlets
- Understanding of agglomerative hierarchical clustering
- General knowledge of interactomics
- Basic competency with search engines
- Writing and presentation skills
- Citation management systems
- Document preparation software
- Coding scripts for data processing and analysis
- Knowledge and understanding of relevant libraries or software packages

5 Milestones

This section outlines some of the targets I set out to accomplish, and comment on when and how they were accomplished.

5.1 Planning and Goals

Before beginning the project it was important to decide what the vision of this project was (see the objective section) and plan how that vision was going to be accomplished. At the level of deadlines, I worked on the parts of the project in a *first-in-first-out* (FIFO) scheduling approach. At the level of the parts of the project, I planned in terms of the dependencies of different steps of the data processing and analysis. For example, I worked on acquiring the data before implementing a way to load it into memory.

5.2 Obtaining Data

One of the most important parts of this project was ensuring I got the data to perform the analysis, so one of the things I checked even before proposing the idea of this project was to ensure that there was an accessible dataset. I found that the entire BioGrid database was open to the public as a tab-delimited file of column-wise panel data, and therefore parsing the data did not involve coding a customize parser.[6] Doing this way back in September 2020, I knew from early on that this project wouldn't have any pitfalls in obtaining a dataset that could be analyzed.

5.3 Running Analysis

With the data in hand, it was necessary to choose the specific implementations of each mathematical procedure and run them at the appropriate state of the data processing pipeline. The first stage was easily completed by using Pandas dataframes, and then split extract the chosen subsets of the data for analysis.[20] Next was the enumeration of the automorphism orbits for each species, which was done with a combination of RAGE[21] and a Python script I coded using NetworkX.[22] With a graphlet orbit matrix for each species, different distance metrics were applied to them in a pairwise fashion to produce a distance matrix, and that distance matrix was then used to hierarchically

[6] Writing parsers for poorly specified file formats can be extremely time-consuming.

cluster the species using an agglomerative strategy.[23] The clustering produced a tree, which was then compared to a reference taxonomy using a confusion matrix method of evaluation on conceivable vs actual partial orders of species.

6 Challenges

This section will discuss some of the technical challenges I encountered during the project that others should anticipate if they attempt to replicate this work.

6.1 Data Coverage and Depth

The first challenge I encountered was that some of the species had far fewer entries than others, which could could imply that sampling errors or publication biases could affect the interspecies comparisons. To avoid this, I filtered the entries by which species had the largest number of entries and kept the top 9 of such species that had over 5 000 entries. This meant of 9/71 reduction in the coverage of species, but a marked increase in the depth of information [7] about each species within the analysis.

6.2 Interaction Directionality

Originally I had intended to used automorphism orbits of *directed* graphlets, implying that the network encoded asymmetric relations. From an early reading of the documentation on the BioGrid dataset, I had believed that the database gave information about directional relations via examples given about gene regulation. However, after contacting the maintainers of the database I learned that in general the directionality of the interactoins was not specified and would not be extractable without manually curating each interaction. Without an asymmetric relation, I decided to perform the analysis on simple graphlets because they would not require an assumption of directionality. This lost information about the interactions in principle, but in practice it allowed for the analysis to be performed.

7 Methodology

In this project I planned my work around the deadlines, and gave myself about a week before each deadline to start working on whatever the required deliverables were.[8] For the most part, I found that this approach was suitable for this project because I already had a pretty clear picture of what I wanted to do from the start. If it were a project where I was much less familiar the general concepts, I would have given myself more time to find out what I didn't know that I didn't know. The total human hours spent on this project was probably around 15-20 hours, including coding, writing and preparing presentations. The graphlet enumerations took up to 8 hours to complete for all species, given that I left the scripts to run overnight and they were done the morning after.

8 Deliverables

Unlike projects that attempt to implement algorithms or build graphical user interfaces where the deliverables should be software packages, this project was scientific in nature in its attempt to

[7] About 1.25 million entries remained after filtering

[8] See the course outline for dates on when each deliverable was due.

extract knowledge from empirical data. Therefore the main deliverable of this project is the answer to the research question, which should be interpreted carefully. See the Results and Discussion section for more information.

9 Learning Outcomes

In addition to the results of this project, the following list are some areas where I improved during the semester while working on this project.

- Improved understanding of graphlets
- Improved understanding of comparing tree structures
- Expand knowledge of Python libraries
- Learn to use RAGE to calculate automorphism orbits
- Practice communicating results in written and verbal form

10 Conclusion

Overall, it does not seem like the graphlet orbit matrix with a distance-based approach was capable of inferring accurate phylogenetic trees. There are other approaches that may be applied to interaction networks that could reveal such information, which are left for future studies.

References

[1] Charles Darwin. *On the Origin of Species by Means of Natural Selection.* Murray, London, 1859. or the Preservation of Favored Races in the Struggle for Life.

[2] Hanne Andersen and Brian Hepburn. Scientific Method. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, winter 2020 edition, 2020.

[3] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond. BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4):e1003537, April 2014.

[4] Howard Ochman, Jeffrey G. Lawrence, and Eduardo A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, May 2000.

[5] Calum Johnston, Bernard Martin, Gwennaele Fichant, Patrice Polard, and Jean-Pierre Claverys. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology*, 12(3):181–196, February 2014.

[6] Anthony Griffiths. *An introduction to genetic analysis.* W.H. Freeman, New York, 2000.

[7] Norton D. Zinder and Joshua Lederberg. Genetic exchange in salmonella. *Journal of Bacteriology*, 64(5):679–699, 1952.

[8] Lars Kiemer and Gianni Cesareni. Comparative interactomics: comparing apples and pears? *Trends in Biotechnology*, 25(10):448–454, October 2007.

[9] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, Frederick Zhang, Sonam Dolma, Andrew Willems, Jasmin Coulombe-Huntington, Andrew Chatr-aryamontri, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, 47(D1):D529–D541, November 2018.

[10] Leonard Euler. Solutio problematis ad geometriam situs pertinentis. pages 128–40, 1736.

[11] Ine Melckenbeeck, Pieter Audenaert, Didier Colle, and Mario Pickavet. Efficiently counting all orbits of graphlets of any order in a graph using autogenerated equations. *Bioinformatics*, 34(8):1372–1380, November 2017.

[12] Anida Sarajlić, Noël Malod-Dognin, Ömer Nebil Yaveroğlu, and Nataša Pržulj. Graphlet-based characterization of directed networks. *Scientific Reports*, 6(1), October 2016.

[13] Yong Zhang, Shaojuan Li, Geir Skogerbø, Zhihua Zhang, Xiaopeng Zhu, Zefeng Zhang, Shiwei Sun, Hongchao Lu, Baochen Shi, and Runsheng Chen. *BMC Bioinformatics*, 7(1):252, 2006.

[14] M. Heymans and A. K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19(Suppl 1):i138–i146, July 2003.

[15] Linjing Yang, Karunesh Arora, William A. Beard, Samuel H. Wilson, and Tamar Schlick. Critical role of magnesium ions in DNA polymerase β's closing and active site assembly. *Journal of the American Chemical Society*, 126(27):8441–8453, July 2004.

[16] Tomaž Hočevar and Janez Demšar. A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4):559–565, February 2014.

[17] B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, October 1975.

[18] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87, January 1922.

[19] Maurice Roux. A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, 35(2):345–366, July 2018.

[20] Wes Mckinney. pandas: a foundational python library for data analysis and statistics. *Python High Performance Science Computer*, 01 2011.

[21] D. Marcus and Y. Shavitt. RAGE: A rapid graphlet enumerator for large networks. *Computer Networks*, 56(2):810–819, February 2012.

[22] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[23] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stefan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, lhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antonio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 2020.