Tree Shape Statistics and their Applications

by

Maryam Hayati

M.Sc., Sharif University of Technology, 2013
B.Sc., University of Tehran, 2007

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY
Spring 2020

Approval

Name:	**Maryam Hayati**
Degree:	**Doctor of Philosophy (Computer Science)**
Title:	**Tree Shape Statistics and their Applications**
Examining Committee:	**Chair:** Igor Shinkar Assistant Professor
	Leonid Chindelevitch Senior Supervisor Assistant Professor
	Caroline Colijn Supervisor Associate Professor
	Maxwell Libbrecht Internal Examiner Assistant Professor School of Computing Science
	Daniel G. Brown External Examiner Professor University of Waterloo, David R. Cheriton School of Computer Science
Date Defended:	**November 29, 2019**

Abstract

Phylogenetic trees are frequently used in biology to study the relationships among several species or organisms. The shape of a phylogenetic tree contains useful information about patterns of speciation and extinction, so powerful tools are needed to investigate the shape of a phylogenetic tree. Tree shape statistics are a common approach to quantify the shape of a phylogenetic tree by encoding it with a single number. Tree shape statistics such as the Sackin and Colless indices have been used in a variety of contexts. Their applications start from differentiating between trees conforming to different stochastic evolution models to more recent developments in phylodynamics, where the tree structures are used to predict short-term growth and fitness. In contrast to the vast application range of tree shape statistics, existing statistics often do not suffice to distinguish between essential scenarios such as trees corresponding to different viral pathogens or different geographical scales for the same pathogen.

In this dissertation, we study tree shape statistics from different aspects. First, we propose a new resolution function to evaluate the power of different tree shape statistics to distinguish between dissimilar trees. Second, we propose two classes of new tree shape statistics. For the first one, we use network science, a well-developed branch of data science, to inspire five novel tree shape statistics. For the second one, we introduce a linear combination of two existing statistics that are optimal with respect to a resolution function and show evidence that the statistics in this class converge to a limiting linear combination as the size of the tree increases. Lastly, we investigate the problem of using tree shape statistics and machine learning tools applied to phylogenetic trees to predict the success of individual influenza virus subtrees.

Furthermore, we study the distribution of the Robinson-Foulds metric. We modify the dynamic programming algorithm for computing the distribution of the Robinson-Foulds distance for a given tree by leveraging the Number-Theoretic Transform (NTT), and improve the running time from $O(n^5)$ to $O(n^3 \log n)$, where n is the number of tips of the tree.

Keywords: Phylogeny; Tree Shape Statistic; Resolution; Network Science; Influenza; Machine Learning; Prediction

Dedication

To my beloved parents Ameneh and Ali who never gave up on me
To my love Ehsan who always supports me

Acknowledgements

First and foremost, I would like to express my special appreciation and thanks to my senior supervisor Dr. Leonid Chindelevitch. It was a great experience working with Leonid. I appreciate all his contribution of time, ideas, patience, and understanding. I would never have been able to finish my dissertation without his guidance and encouragement. I would like to sincerely thank my supervisor Dr. Caroline Colijn for her support, guidance, and encouragement. Not only did I learn a lot of things in bioinformatics from her, but she also taught me invaluable life lessons. I would also like to thank Dr. Daniel G. Brown and Dr. Maxwell Libbrecht for taking the time and effort to read my thesis and giving me very valuable comments.

I am forever indebted to all my co-authors, the results in this thesis were based on a joint work with DR. Bita Shadgar, Dr. Priscila Biller, and Dr. Art Poon. I thank them for all the things they taught me. I would like to acknowledge the administrative and technical staff in the School of Computing Science for making this school a productive environment.

My friends at Simon Fraser University make my studies more pleasant and enjoyable. Especially, I am grateful to Golnaz Gharachorlu, Pinar kavak, Rashme Acharya, Sahand Khakabimamaghani, Azadeh Zamani, Maryam Razmhosseini, Zahra Zohrehvand, Nafiseh Sedaghat, Faezeh Bayat, and Sheida Alen. I would also like to specially thank my lovely friends Zahra Nazari, Bahareh Najmi, Hoda Heidary, Fatemeh Sagharchi, Somayeh Razaghi, and Shaghayegh Farzaneh.

Heartfelt thanks goes to my parents and my wonderful siblings: Alireza, Azam, Saeid, and Elaheh. I appreciate their endless love and support that made these years a pleasant chapter of my life. Words cannot express how grateful I am to my parents for all of the sacrifices that they have made on my behalf. They are always my sense of safety and calmness. Thank you for giving me everything in life and for being so loving, caring and supportive.

Last but not least, I am immensely grateful to my husband, Ehsan, for all his love, patience and kindness. He has been there from the very start of this journey as an influential mentor. I could not have completed this work without his unfaltering love and support.

Previous Publications of This Work

Chapter 3 and portions of chapter 4 are based on the paper “A new resolution function to evaluate tree shape statistics” published in PLOS One, co-authored with Bita Shadgar and Leonid Chindelevitch. They are reprinted with the kind permission of the editors.

Portions of Chapter 4 is based on the paper “Network science inspires novel tree shape statistics” currently in bioRxiv [30] and submitted to PLOS Computational Biology. It is included with the kind permission of Leonid Chindelevitch, Art Poon and Caroline Colijn.

Chapter 5 is based on the paper “Predicting the short-term success of human influenza virus variants with machine learning” currently on bioRxiv [74] and submitted to the Proceeding of the Royal Society B co-authored with Priscila Biller and Caroline Colijn. It is reprinted with the kind permission of the authors.

Chapter 6 is based on the paper “Computing the distribution of the Robinson-Foulds distance” accepted in APBC 2020.

Contents

Approval ii

Abstract iii

Dedication iv

Acknowledgements v

Table of Contents vii

List of Tables x

List of Figures xii

1 Introduction 1

1.1 Preliminaries 1

1.2 Evolutionary Models of Phylogenetic Trees 3

1.2.1 Equal-rates Markov Model 3

1.2.2 Coalescent Model 4

1.2.3 Proportional-to-Distinguishable-Arrangements Model 4

1.2.4 Equiprobable Model 5

1.2.5 Beta-splitting Family of Distributions on Cladograms 5

1.3 Tree Comparison 6

1.3.1 Nearest Neighbor Interchange Metric 6

1.3.2 Subtree Prune Regraft Metric 6

1.3.3 Robinson-Foulds Metric 7

1.3.4 Quartet Metric 8

1.4 Phylogenetic Reconstruction Methods 8

1.4.1 Distance-Based Methods 9

1.4.2 Character-Based Methods 9

1.5 Tree Shape Statistic 11

1.6 Organization of this Thesis 12

2 Literature Review 14

2.1 Evaluating the Power of Tree Shape Statistics . . . 14

2.2 Tree Shape Statistics . . . 16

2.2.1 Classical Tree Shape Statistics . . . 16

2.2.2 Recently Proposed Tree Shape Statistics . . . 24

2.3 Applications of Tree Shape Statistics . . . 29

2.3.1 Testing Evolutionary Hypotheses . . . 29

2.3.2 Applications in Phylodynamics . . . 30

2.4 Computing the Distribution of the Robinson-Foulds Distance . . . 33

3 Evaluating the Power of Tree Shape Statistics 35

3.1 Problem Definition . . . 36

3.2 Methods . . . 36

3.2.1 Geometric Approach . . . 36

3.2.2 NNI Metric for Unlabeled Trees . . . 37

3.2.3 Resolution of Tree Shape Statistics Based on a Distance Matrix . . . 39

3.2.4 Resolution of Tree Shape Statistics Based on the Laplacian Matrix . 42

3.2.5 The Upper and Lower Bounds . . . 43

3.3 Results . . . 44

4 New Tree Shape Statistics 47

4.1 Problem Definition . . . 49

4.2 Methods . . . 49

4.2.1 Linear Combination of the Classical Tree Shape Statistics . . . 49

4.2.2 Tree shape Summaries Based on Network Science . . . 50

4.2.3 Spectral Properties Derived from the Distance Laplacian Matrix . . 53

4.2.4 Data and Simulations . . . 53

4.3 Results . . . 56

4.3.1 Comparing the Power of Saless with the Classical Statistics . . . 56

4.3.2 Linear Combination of the Classical Tree Shape Statistics . . . 56

4.3.3 Network Statistics Improve Classification of the Trees Reconstructed From Different Viruses and Epidemiological Scenarios . . . 57

5 An Application of Tree Shape Statistics 61

5.1 Problem Definition . . . 62

5.2 Methods . . . 63

5.2.1 Reconstructing Influenza Virus Trees . . . 63

5.2.2 Subtree Extraction . . . 63

5.2.3 Features . . . 65

5.2.4 Success and Training Approach . . . 68

5.2.5 Classification 68
5.2.6 Time Slices 69
5.3 Results 71
5.3.1 Cross-Validation on Different Experiments 74
5.3.2 Results of Time Slices 76
5.3.3 Model Parameters 77
5.3.4 Influenza Virus B 79
5.3.5 Best Features 79
5.3.6 Predictions on the Recent Subtrees 80

6 Computing the Distribution of the Robinson-Foulds Distance 84
6.1 Problem Definition 85
6.2 Methods 85
6.2.1 Computing the Distribution of the RF Distance 86
6.2.2 Computing the Distribution of the RF Distance with the Fast Fourier Transform 89
6.2.3 Computing the Distribution of the RF Distance with the Number-Theoretic Transform 90
6.3 Results 92

7 Conclusions 96
7.1 Summary 96
7.2 Discussion 98
7.3 Future Work 101

Bibliography 103

Appendix A Results Corresponding to Chapter 4 117

Appendix B Results Corresponding to Chapter 5 123

List of Tables

Table 2.1 Brief definition for tree shape statistics. Here r_i and s_i are the number of tips of the left and right subtrees of an internal node respectively. n is the number of tips of a subtree. n_i is the number of nodes at depth i, M_i represents the height of the subtree rooted at an internal node i, and N_i is equal to the depth of tip i. A ladder in a tree is a set of consecutive nodes with one tip child. We represent the set of all internal nodes of a tree by $\mathcal{I}$, the set of all tips (or external nodes) by $\mathcal{L}$. The tree shape statistics induced by betweenness centrality, closeness centrality and eigenvector centrality are defined as the maximum values of each centrality over all the nodes of a tree, and distances are in units of number of edges (without branch lengths). We introduce the network statistics in chapter 4 of this thesis. 28

Table 3.1 Scaled resolution scores for the classical tree shape statistics based on our resolution function. n is the number of leaves. The best classical statistic is the Colless index and the worst ones are B_1 and I_2 (the same ranking as for the previous resolution function). The highlighted values correspond to the best statistics in each row. 45

Table 4.1 The value of λ and the resulting resolution R_D for trees with different number of leaves. 50

Table 4.2 Scaled resolution scores for tree statistics based on the distance resolution function on the *NNI* distance matrix. The resolution is between 0 and 1. n is the number of leaves. The tree shape statistics are described in chapter 2. The highlighted values correspond to the best statistics in each row. 56

Table 4.3 Scaled resolution scores for the optimal linear combinations of $I - B_2$, B_2-B_1, and *Saless* based on our new proposed resolution function (Laplacian resolution function). The corresponding optimal values of λ are shown next to each combination. 57

Table 4.4 Summary measures for phylogenetic trees. Here, n_i is the number of nodes at depth i. 58

Table 5.1 Brief definition for tree shape statistics. Here r_i and s_i are the number of tips of the left and right subtrees of an internal node respectively. n is the number of tips of a subtree. n_i is the number of nodes at depth i, M_i represents the height of the subtree rooted at an internal node i, and N_i is equal to the depth of node i. A ladder in a tree is a set of consecutive nodes with one tip child. We represent the set of all internal nodes of a tree by $\mathcal{I}$, the set of all tips (or external nodes) by $\mathcal{L}$. In "generalized branching next" we chose $m = 2$. Skewness and Kurtosis are two measures to describe the degree of asymmetry of a distribution [114]. The tree shape statistics induced by betweenness centrality, closeness centrality and eigenvector centrality are defined as the maximum values of each centrality over all the nodes of a tree, and distances are in units of number of edges (without branch lengths). Features called "natural" may not have been used as tree features previously but are natural extensions of simple features (eg skewness is a natural quantity to compute). The network science properties were computed in R using the treeCentrality package [101] and the tree-wide summaries were primarily obtained using the phyloTop package [91] . 67

Table 5.2 Our subtree selection algorithm uses three parameters: a minimum subtree size, a maximum allowable overlap and the length of the time window (Figure 6.1). We denote our default as "ALT0" and our alternatives as "ALT1" and "ALT2". There are natural trade-offs: a larger minimum size, lower overlap and longer time frame all result in fewer accepted subtrees. We found good performance for each of these three alternate setups. 78

Table A.1 Accession numbers for outgroups 118

List of Figures

Figure 1.1 (a) shows the caterpillar (completely asymmetric tree or completely unbalanced tree) on 7 leaves, and (b) shows the maximally balanced tree (completely symmetric tree or completely balanced tree) on 7 leaves. 3

Figure 1.2 There are three possible options to convert a rooted labeled phylogenetic tree with 3 tips to a tree with 4 tips through speciation of one lineage. The two left trees have the same shape [124]. 4

Figure 1.3 Two possible configurations for an NNI move on a rooted tree. . . . 7

Figure 1.4 This figure shows two binary unrooted phylogenetic trees with 5 tips. The Robinson-Foulds distance between these two trees is 4. 8

Figure 1.5 The three possible quartet topologies of species A, B, C, and D. . . 8

Figure 2.1 Log-log scale plot of the splits of a tree from a real dataset, and approximate median lines for the beta splitting model[6]. 27

Figure 3.1 The geometric perspective of a good and bad tree shape statistic. From a geometric perspective, a good statistic can discriminate between different trees, and place similar trees together. In theses two figures, we embedded the set of trees with 9 tips using multidimensional scaling (*MDS*) and the *NNI* distance between the trees. The points in the top and bottom plots are colored based on their Sackin and I2 values respectively. The green, blue and red points correspond to the upper quartile, lower quartile, and the inter-quartile interval of the distribution of the statistics, respectively. The clustering pattern in the top figure indicates that the Sackin index can separate the trees into groups in a way consistent with the *NNI* distances, while the I2 index is unable to do so. This can also be seen in Table 3.1 using the resolution function for these two statistics. 38

Figure 3.2 This figure shows that the initial random tip labels assigned by the *nni* function would not affect the *NNI* distance between two unlabeled trees. Consider a single tree shape with two sets of tip labels. Each column shows the set of all labeled rooted trees that are at distance one from the corresponding tree. Each tree shape is assigned a unique number using a numbering scheme. Since the numbering scheme ignores the tip labels, the assigned numbers in these two sets are exactly the same. The set of all unique unlabeled rooted trees which are at distance one from the given tree is computed by removing the duplicates from the set of numbers. 40

Figure 3.3 This figure depicts the definition of unlabeled *NNI* metric. Each node is a binary unlabeled rooted tree with 6 tips. Two trees are connected with an edge if they are one single *NNI* move apart. (Figure is taken from [115]) 41

Figure 3.4 This plot shows the running time of computing the Laplacian matrix (dashed line) and the distance matrix (solid line) as well as the growth of the number of trees by increasing the number of leaves (dot line). To compute the geometric resolution one must compute the distance matrix; however, to compute our proposed resolution one needs only compute the Laplacian matrix. The left axis shows the time on a log scale. The bottom axis shows the number of leaves, and the right axis shows the number of trees with a specific number of leaves. The base unit for time is the second and a base-10 log scale is used for the left and right axis. A comparison between the slopes of the dashed line and the solid line in the plot shows that the running time of our proposed resolution is much faster that that of the previous method. The computational experiments were conducted on an Intel Core i7 with 2.4 GHz 64-bit processor, 16.0 GB of RAM and macOS Sierra 64-bit as the Operating System. 46

Figure 4.1 The value of λ appears to converge as the number of leaves grows. . 50

Figure 4.2 A sample phylogenetic tree with the associated centrality values at each node. (a) An example tree. In the epidemiological context, tips ($a - g$) would correspond to pathogen sequences and internal nodes ($A - F$) to their inferred common ancestors. Node D subtends a "cherry" configuration, and node C subtends two cherries (a "double cherry"). The heights of the internal nodes are 1 (E, F), 2 (C, D), 4 (B) and 8 (A), so the diameter is 16 and the Wiener index is 484, for a mean path length of 6.21. (b) Same tree with betweenness centrality values at each node (note that branch lengths do not change them). The tree has betwenness centrality 45. (c) Same tree with farness (reciprocal of closeness centrality) values at each node. The tree has closeness centrality 1/48. (d) Same tree with eigenvector centrality values (scaled to have a minimum of 1) at each node, rounded to 3 significant figures. Here, the leading eigenvalue is $\lambda = 9.05$. By our definition, the tree has eigenvector centrality $714/1023 = 0.698$ (here, 1023^2 is the sum of the squared values). 54

Figure 4.3 (a) Accuracy of classification with and without network science features, as well as with and without the distance Laplacian spectral features. (b) Feature importance in multi-class classification across all scenarios, ordered by the median. Each point is the rank of the corresponding feature in one of the classification tasks. Low ranks correspond to the most important features (i.e. the top-ranked feature has rank 1, and so on). 60

Figure 5.1 Illustration of the formation of trimmed subtrees: (a) the circled clade contains a subtree; (b) red branches reach tips that occur after the trimming time period and so are pruned out; (c) the resulting trimmed subtree. 68

Figure 5.2 This figure illustrates our training and testing approach for experiments on influenza virus H3N2, influenza virus H1N1, influenza virus B and pooling the subtreees . We divide the subtrees whose outcome is known into training (75%) and testing (25%) data and choose hyperparameters using 10-fold cross-validation on the training data only. We use those hyperparameters to train the "general" model and test it on the testing data. (b) The schematic figure for training on a tree (H3N2) and testing on another tree (H1N1). 70

Figure 5.3 This figure depicts the definition of a successful subtree in the time slicing approach. The tree on the left, T_i, represents a tree reconstructed from a set of sequences up to time i and the tree on the right shows the same tree after one year (T_{i+1}). We compare these two trees to predict the successful subtrees in the time slicing approach. For each subtree c we compare the size of the subtree in T_i (X) with the size of the subtree rooted at the most recent common ancestor of the tips of subtree c in tree T_{i+1} (Y), in an interval of 3.4 years following the root of the subtree. If $Y > \alpha X$ ($\alpha \in \{1.1, 1.2, 1.3, 1.5\}$) then we say subtree c is successful. This overcomes the challenge that T_{i+1} does not contain the root of c; it does contain a node that is the MRCA of the tips in subtree c. 71

Figure 5.4 Phylogenetic tree reconstructed from H3N2 subtype sequences using RAxML, with tips highlighted. Each yellow tip is in a trimmed subtree (7615 out of 12785 tips); grey tips are not. The sequences are downloaded from GenBank with dates from 1980 to 2018-5. Long branches in this timed tree did not appear as long branches in the RAxML tree and were not removed (though their tips are not in any trimmed subtree). Inset: illustration of formation of trimmed subtrees: (a) the circled clade contains a subtree; (b) red branches reach tips that occur after the trimming time period and so are pruned out; (c) the resulting trimmed subtree. 72

Figure 5.5 (a) ROC showing performance of linear kernel *SVM* trained and tested on H3N2, trained and tested on H1N1, trained and tested on B, trained and tested on the merged subtrees of H3N2 and H1N1 and trained on H3N2 and tested on H1N1. Figure 5.6 shows variation in these curves and their AUCs over 10 models trained on 90% of the test data for each case. (b) ROC showing classification performance for restricted sets of features. Top: tree shape features (no branch lengths). Epi: epitope features. LBI: local branching index. B.Len: some features include tree branch lengths. 73

Figure 5.6 (a) The result of 10-fold cross validation for *SVM* models trained on the subtrees whose ancestral nodes occurred pre-2015-1 or which began later but whose outcomes are known. AUC values range from 0.73 to 0.90 (average 0.82) with 80% of the folds resulting in more than 0.75 AUC. (b) 10-fold cross-validation on the H1N1 tree using *SVM* with linear kernel and a set of topological properties of the clades. AUC values range from 0.52 to 0.95 (average 0.76) with AUC more than 0.75 in 70% of the folds. (c) The result of 10-fold cross-validation on the merged dataset of H3N2 and H1N1 trees. The minimum and maximum values of AUCs among the folds are 0.75 and 0.88 respectively (average 0.82). (d) 10-fold cross-validation for training on the H3N2 tree and testing on the H1N1 tree; we removed 10% of the training data at each fold and evaluated the model on the test data. AUC values range from 0.72 to 0.75 (average 0.74). . . . 75

Figure 5.7 We classified the subtrees of the influenza virus H3N2 tree reconstructed from the sequences up to time i by comparing with the tree reconstructed from sequences up to time $i+1$ which $i \in \{2012-5, 2013-5, 2014-5, 2015-5, 2016-5, 2017-5\}$. In this figure the result of classification using different ratio between the size of a subtree in time i and the size of the corresponding subtree in time $i+1$ is shown. 76

Figure 5.8 (a) Correlation between the size of trimmed subtrees (x-axis) and the rate of success (y-axis) for the 391 subtrees from the H3N2-HA dataset. The rate of success is defined as the number of succesful trees divided by the total number of trees for a certain range of sizes. The ranges were computed in order to encompass approximately the same number of subtrees, and the color of the bars represent how many subtrees were taken into account for the computation of the success rate. The subtrees vary from size 8 up to 460, but as most of the dataset is composed by small subtrees, we used a log scale to better visualize the information. The subtree sizes highlighted in yellow are detailed in the bottom panel. (b) Subtree size distribution for trees up to size 50, which corresponds to 87.4% of the dataset. For each size, the graph shows how many subtrees were succesful (blue blar) and unsuccesful (red bar). 77

Figure 5.9 The ROC curves from prediction for the H3N2-HA dataset using different models and different definitions of a successful subtree. In all models, setting the threshold to an appropriate value allows a balance between successful and unsuccessful outcomes, resulting in better performance. Among different sets of parameters for extracting the subtrees, ALT0 is the most powerful model resulting in 0.89 AUC. In all of these experiments we used the topological and epitope features. 78

Figure 5.10 We find that a choice of $\alpha = 1.2$ results in a good balance between successful and unsuccessful subtrees. Combining H1N1 and B subtrees results in good performance, suggesting similarities between the link between subtree structure and success between H1N1 and influenza virus B. 79

Figure 5.11 This figure shows the importance of each feature in the classification task. We use an ensemble technique to find the importance of each feature. MeanEp, MaxEp, and MeadianEp are the three features we defined on the epitope sites of the sequences (see Section 5.2.3 for more details). numberTipsTrimmed is the number of tips in the trimmed subtrees. The rest of the features are defined in the Table 5.1. 80

Figure 5.12 For the prediction task, we divide our subtrees into recent (subtrees originating after 2015-1) and non-recent sets (the rest of the subtrees). We then trained 10 models using 10-fold cross validation together with one general model on the set of non-recent subtrees and used these 11 models for prediction on the recent subtrees. 81

Figure 5.13 Summary of predictions for the recent subtrees (for detailed information, see Figure 5.14). (a) Relations between the clades defined by the World Health Organisation (WHO), showing the emergence of new clade designations from existing ones; (b) Frequency of clades through years, based on 3298 H3N2 HA sequences sampled between 2015 and 2018. Every sequence is associated with a tip of one or more subtrees used in the predictions. The clade designation of the tips determines the clade designation of the subtrees by majority rule. (c) Predictions for the recent subtrees. Among the 120 recent subtrees, the outcome (success/failure) of 63 subtrees is not known. Each subtree was tested on 11 different SVM models (see Figure 5.12). A rectangle corresponds to the prediction of a clade, and its area is proportional to the number of subtrees used in the prediction. The number of subtrees associated with a clade is indicated in parentheses. Color reflect the combined predictions of the subtrees associated with each clade. 82

Figure 5.14 Predictions of the rate of success for the clades defined by the World Health Organisation (WHO). A column contains the information about one clade, indicated on the top and following the same color code as Figure 5.13. The columns (clades) are sorted according to the rate of success, from the most successful (left) to the least successful (right). The rate of success of a clade is an average of all predictions from subtrees whose majority of tips contain DNA markers linked to the clade (11 predictions per tree). All subtrees used in these predictions were extracted between 2015 and 2018 and, thus, have an unknown outcome, as the actual growth rate in 3.4 years from the initial node is still inaccessible. The subtrees are classified in four types: (A) subtrees whose prediction is to succeed but the current growth is still below the threshold considered to be successful; (B) subtrees whose prediction is to fail and the current growth is below the threshold considered to be successful; (C) subtrees whose prediction is to succeed and the current growth is above the threshold considered to be successful (notice that even in this scenario it is not possible to be certain, since the structure of the tree may change); (D) subtrees whose prediction is to fail but the current growth is above the threshold considered to be successful. Schematics for each case are presented on the bottom of the table. 83

Figure 6.1 There are three types of internal nodes in a binary rooted tree. X is a type I node, i.e. an internal node with two tip descendants (cherry). Y is a type II internal node since it has one tip and one internal node descendants. Z is a type III internal node since it has two internal node descendants. 86

Figure 6.2 This figure is an example of how to compute $R(v, s, k)$. Consider the subtree rooted at v. For $k = 1$ and $s = 2$ there are four options for E: $E_1 = \{e_2, e_4\}$, $E_2 = \{e_3, e_4\}$, $E_3 = \{e_1, e_2\}$, and $E_4 = \{e_1, e_3\}$. The value of $N_E(T_v)$ for each of these sets is $\mathcal{B}(0)\mathcal{B}(1)\mathcal{B}(0)$, $\mathcal{B}(1)\mathcal{B}(0)\mathcal{B}(0)$, $\mathcal{B}(0)\mathcal{B}(1)\mathcal{B}(1)$ and $\mathcal{B}(1)\mathcal{B}(0)\mathcal{B}(1)$ respectively, so $R(v, s, k) = 24$ . . 88

Figure 6.3 (a) shows the maximally balanced (completely balanced unrooted tree) with 8 tips, and (b) shows the unrooted caterpillar (completely unbalanced tree) on 8 tips 93

Figure 6.4 Log scale plot showing the RF distribution of the completely balanced and the completely unbalanced labeled trees with $8, 16, 32$ and 64 tips. The y-axis represents the frequency of the distances from a given phylogenetic tree. 94

Figure 6.5 a) log-log plot showing the running time of the algorithm when we do not use the NTT to compute the convolution. The slope of this plot is 5.035, in agreement with the $O(n^5)$ running time of the algorithm. b) log-log plot showing the running time of the algorithm when we use NTT to compute the convolution. There are some jumps in the plot, indicating the points where the transformation length changed. This figure shows that using NTT to compute the convolution can improve the performance of the computation of the RF distance distribution. 95

Figure 6.6 log-log plot showing the running time of computing the convolution using NTT vs. transformation length. The running time changes nearly linearly with the transformation length. 95

Figure A.1 Tree summaries for the HIV/Dengue/Measles, influenza and simulations of size 100. 119

Figure A.2 Tree summaries for HIV in three settings and simulations of size 300 120

Figure A.3 A randomly sampled tree from each scenario (except HIV in the 3-virus comparison because HIV is represented in three other trees). To allow for focus on tree shape rather than on branch lengths, trees have been visualized with branch lengths set to 1. 121

Figure B.1 Distribution of H3N2-HA sequences between 1980 and 2018. The blue bar corresponds to sequences assigned to one or more clades, whereas the red bar corresponds to sequences without clade designation. The rate of unassigned cases drops considerably for recent sequences (from 2011 on), which encompass most of the analyzed dataset (73.45%). 123

Figure B.2 Correlation between Local Branching Index (LBI) and Katz Centrality measures (weighted version, where the distance between two nodes is defined as the sum of the branch lengths in the path separating them). (a) Each point corresponds to the LBI and Katz Centrality measures computed for the root of the 391 subtrees from the H3N2-HA dataset. Parameters used: $\tau = 50$ (LBI), $\alpha = 0.95$ (Katz Centrality). (b) Correlation for 10 trees simulated with a pure birth process. All trees have 500 leaves, and each point corresponds to the LBI and Katz Centrality measures computed for a node of those trees. Parameters used: $\tau = 10$ (LBI), $\alpha = 0.95$ (Katz Centrality). 124

Chapter 1

Introduction

With the exponential growth of genome databases, the importance of phylogenetics has increased dramatically over the past years. Studying phylogenetic trees enables us not only to understand how genes, genomes, and species evolve but also helps us predict how they might change in the future. This thesis studies phylogenetic trees with a focus on the shape of the trees.

In this chapter, we introduce phylogenetic trees and some terminologies related to them, including evolutionary models of phylogenetic trees, tree comparison metrics, and tree shape statistics.

1.1 Preliminaries

A phylogenetic tree is a connected acyclic graph $T = (V, E)$ with vertex set V $(V(T))$ and edge set E $(E(T))$. Given a phylogenetic tree T, a leaf (also called an external node or tip) of T is a node of degree one. An internal node of T is any non-leaf node of the tree. Herein, $\mathcal{I}$ represents the set of all internal nodes of a tree and $\mathcal{L}$ denotes the set of all leaves (or external nodes). A phylogenetic tree is called a labeled phylogenetic tree if the tip labels are considered; otherwise, it is called an unlabeled phylogenetic tree. A phylogenetic tree is weighted if there is a function $w : E \to \mathbb{R}$ that assigns a weight $w(uv)$ to each edge $uv \in E$. The weight $w(e)$ of an edge $e = uv$ of a tree is also called its branch length. Throughout this thesis, we do not consider the branch lengths of a phylogenetic tree except in Chapter 5, which the branch lengths represent the time distance between the branching events. The shape of a phylogenetic tree is the structure or topology of a phylogenetic tree without considering tip labels and branch lengths. The shape of a phylogeny tells the story of an evolutionary history going back through time to the most recent common ancestor at the root of the tree.

A phylogenetic tree can be rooted or unrooted. A rooted phylogenetic tree is a tree in which a particular internal node called the root is distinguished from the others; it is postulated to be the ancestor of all the other nodes in the tree. In a rooted phylogenetic tree T, the parent of a node i is the node following it on the unique path from the node to the root r of T; all nodes of T except its root r have a parent. A child of a node i is any node whose parent is i.

Given a node i of T, an ancestor node of i is a node on the unique path from i to the root of T. The descendants of i are all the nodes of T that have i as an ancestor node. In a rooted phylogenetic tree T, the set of all descendants of a node u including u itself forms a subtree; we say that u subtends this subtree.

A rooted phylogenetic tree is binary (bifurcating) if all its internal nodes have exactly two children. In this thesis, we mostly consider rooted binary trees with n leaves, and since then, by a phylogenetic tree (or simply a tree) T, we mean a binary tree (it can be rooted or unrooted and labeled or unlabeled and we emphasis on that whenever is needed). One can easily prove that a rooted phylogenetic tree with n leaves has exactly $(n-1)$ internal nodes and a total of $N = 2n-1$ nodes [50]. The number n_T of unlabeled rooted phylogenetic trees on n leaves grows exponentially with n – asymptotically, $n_T \sim b^n n^{-3/2}$, where $b \approx 2.483$ [115]. The number of labeled rooted and labeled unrooted phylogenetic trees with n tips are $T_n = (2n-3)!!$ and $b(n) = (2n-5)!!$ respectively. Briefly, throughout this thesis, we use these different type of trees and we mention which type of trees we would use at the beginning of each section .

- Phylogenetic tree (or tree): a binary tree that can be rooted or unrooted and labeled or unlabeled.
- Unlabeled phylogenetic tree: a binary tree without tip labels.
- Labeled phylogenetic tree: a binary tree with tip labels.
- Unlabeled rooted phylogenetic tree: a rooted binary tree without tip labels.
- Unrooted labeled phylogenetic tree: an unrooted binary tree with tip labels.

The depth of a node i in a rooted phylogenetic tree T is the number of edges on the unique path from the root of T to i; the root is the only node at depth 0. The height of i is the number of edges on the longest path from i to a leaf of the subtree rooted at i. The height of a rooted phylogenetic tree is the height of its root. The subtree of T rooted at i is the tree induced by i and all of its descendants in T.

A rooted caterpillar (or the completely asymmetric tree) is the unique rooted phylogenetic tree T such that all the internal nodes of T have a leaf child, see Figure 1.1 (a). If i is an internal node in a rooted phylogenetic tree T, the balance value of i is $bal_T(i) = |r_i - s_i|$, where r_i and s_i are the number of tips in the left and right subtrees of the subtree rooted at i. An internal node of T is balanced if $bal_T(i) \leq 1$. T is maximally balanced (completely symmetric) if all of its internal nodes are balanced, and there is a unique maximally balanced phylogenetic tree with n leaves, up to isomorphism, see Figure 1.1 (b).

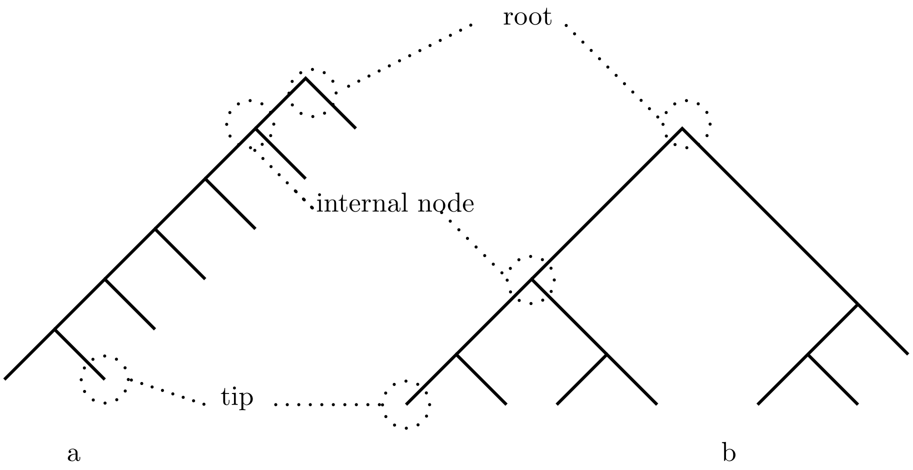


Figure 1.1: (a) shows the caterpillar (completely asymmetric tree or completely unbalanced tree) on 7 leaves, and (b) shows the maximally balanced tree (completely symmetric tree or completely balanced tree) on 7 leaves.

1.2 Evolutionary Models of Phylogenetic Trees

One of the most critical problems in population and evolutionary biology is to find some mathematically simple and biologically plausible stochastic models for rooted phylogenetic trees [124, 6]. Macroevolutionary hypotheses are commonly tested by comparing the shape indices of inferred phylogenetic trees from real data with those predicted by the null models [201, 176, 93]. Several null models have been proposed for phylogenetic trees: the equal rate Markov model (*ERM or Yule*), the proportional-to-distinguishable arrangements model (*PDA*), the Aldous branching (*AB*) model, and the equiprobable model. These models predict the distribution of tree shapes and can be used for hypothesis testing of an estimated phylogenetic tree [124]. The following sections describe the null models mentioned above in detail.

1.2.1 Equal-rates Markov Model

The Equal-rates Markov Model [124] is one of the simplest and most-often postulated among the null models for phylogenetic tree shapes. This model is based on the diversification process of the *Yule* model [204]. Under this model, each extant lineage has the same probability of speciation in the interval $(t, t + \Delta t)$. The probability of each distinct tree shape with n leaves under the *ERM* model is different. The *ERM* model does not directly consider the extinction rate and is a pure birth process. In order to involve the effect of extinction rate, the rate of diversification (the actual rate of specification minus the rate of extinction) is used in the *ERM* [124]. To illustrate the equal-rates Markov model, consider a rooted labeled phylogenetic tree with 4 tips, which is the result of three speciation events. The last branching event produces three rooted labeled phylogenetic trees with equal probabilities such that two of them have the same shape, so the probabilities of the balanced and imbalanced rooted phylogenetic trees are $\frac{1}{3}$ and $\frac{2}{3}$ respectively, see Figure 1.2. Studies on the estimated rooted labeled phylogenetic trees show that even small estimated rooted labeled phylogenetic trees are significantly more imbalanced than expected under the *ERM* model, which is a proof for the existence of variation in speciation and extinction rates among lineages of a phylogenetic tree [17].

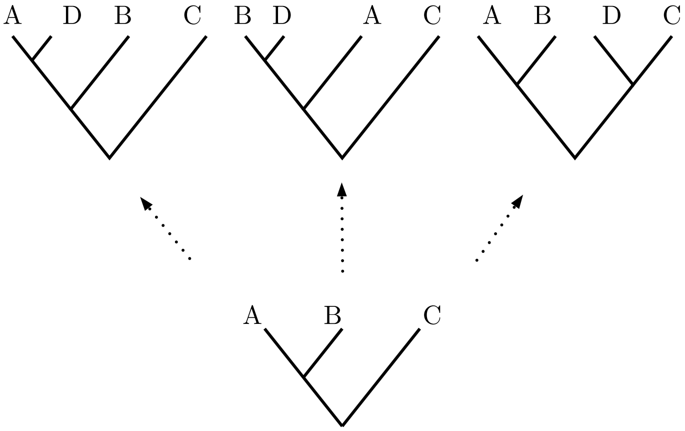


Figure 1.2: There are three possible options to convert a rooted labeled phylogenetic tree with 3 tips to a tree with 4 tips through speciation of one lineage. The two left trees have the same shape [124].

1.2.2 Coalescent Model

Another widely used evolutionary model is the *Coalescent* model [92]. The probability distribution of the coalescent model and the *ERM* model are the same for the shape of a phylogenetic tree. The coalescent model traces the ancestral lineages, which are the series of ancestors of the tips, back through time. A set of n tips comprises $n-1$ coalescent events. Each coalescent event decreases the number of ancestral lineages by one. At each coalescent event, two of the lineages merge into one common-ancestral lineage. Consider n lineages at the present time, and then the number of lineages decreases from n to $n-1$, then from $n-1$ to $n-2$, etc., and finally from two to one through a series of steps. At the final step, there is one single lineage, which is the most recent common ancestor (MRCA) of all tips. [6].

1.2.3 Proportional-to-Distinguishable-Arrangements Model

Under the proportional-to-distinguishable-arrangements model (*PDA*)[124, 159] there is no particular model of growing trees, and each possible rooted labeled phylogenetic tree with n leaves has the same probability. The frequency of each shape (tree shape without considering the tip labels) with n leaves is proportional to the number of different phylogenetic trees that share this phylogeny [124]. For example for a set of phylogenetic trees with 4 leaves, there are two possible shapes; The frequencies of imbalanced and balanced phylogenetic trees under the *PDA* are 0.8 and 0.2, respectively. Therefore, the *PDA* model produces more imbalanced trees than the *ERM* model [124]. For real data, the imbalance of the inferred trees falls somewhere between the value predicted by the *ERM* model and those predicted by the *PDA* model [6, 144].

1.2.4 Equiprobable Model

Under the equiprobable type model *(EPT)*, each possible shape with n leaves has the same probability. For instance, for a set of rooted labeled trees with 4 leaves, the probability of each possible shape is 0.5 [124]. The *EPT* model usually produces rooted trees which are more balanced than *ERM* expectation. There is no plausible model to support the equiprobable model as there is for the Markovian. Therefore, this model is usually discarded in phylogenetics [176].

1.2.5 Beta-splitting Family of Distributions on Cladograms

Both the *ERM* and the *PDA* models can be considered as branching Markov processes [17]. These processes are discrete recursive structures defined by symmetric split distribution. If $P(i|n)$ denotes the probability distribution of the left sister size (i) given the size of the parent clade (n) then under the *ERM* model, the probability that the left sister clade contains i tips is equal to:

$$P(i|n) = \frac{1}{n-1}$$

Under the *PDA* model, the split distribution is defined as follows:

$$P(i|n) = \frac{1}{2}\binom{n}{i}\frac{T_i T_{n-i}}{T_n}, 1 \leq i \leq n-1,$$

where $T_n = (2n-3)!!$ is the number of labeled rooted phylogenetic trees with n tips. Aldous's branching *(AB)* model is defined by the following distribution:

$$P(i|n) = \frac{1}{2h_{n-1}}\frac{n}{i(n-1)}, 1 \leq i \leq n-1,$$

where h_n is the n^{th} harmonic number and is defined as:

$$h_n = \sum_{i=1}^{n}\frac{1}{i}, n \geq 2$$

ERM, *PDA*, and *AB* are particular cases of a one-parameter *beta-splitting* family of distributions on cladograms which is formulated as:

$$P(i|n) = \frac{1}{a_n(\beta)} \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)}, 1 \leq i \leq n-1,$$

where $\Gamma(z)$ is the Gamma function [3] and $a_n(\beta)$ is a normalizing factor [5]. β varies over the range $-2 < \beta \leq \infty$, and *ERM*, *PDA*, and *AB* correspond to $\beta = 0$, $\beta = -1.5$, and $\beta = -1$ respectively and a larger β leads to a more balanced rooted tree [5, 6, 17].

1.3 Tree Comparison

One of the most important problems in phylogenetics is to compute the distance between two phylogenetic trees. Tree comparisons are used for different purposes, ranging from checking the consistency between different tree reconstruction algorithms to deciphering evolutionary associations among organisms and geographical areas [45].

Different methods have been proposed to quantify the similarity between two topologies [27, 40, 41, 84, 85, 156, 157, 178]. These methods can be broadly categorized into two different groups. The first type of metrics counts the minimum number of operations required to transform T_1 into T_2. Two well-known distances in this group are the *nearest neighbor interchange (NNI)* [41, 42, 97, 107] and *subtree-prune-regraft (SPR)* [178] distances, which both are *NP*-hard to compute [77, 97]. The second type of distances represents the two trees as the sets of simpler structures, such as bipartitions, quartets, or clusters, and then computes various measures of similarity between the sets. The *Robinson-Foulds (RF) distance* [157] and the *Quartet distance* [27] are two distances from this group. These metrics can be computed in $O(n)$ and $O(n \log^2 n)$ time, respectively, where n is the number of tips. Thus, the *RF* distance can be computed quickly, which makes it especially popular.

1.3.1 Nearest Neighbor Interchange Metric

A single *NNI* operation in a labeled phylogenetic tree swaps two subtrees that are separated by an internal edge (an edge is internal if neither of its endpoints is a leaf). The two possible *NNI* moves are depicted in Fig 1.3. A labeled phylogenetic tree with n leaves has $O(n)$ neighbors that can be obtained from it via an *NNI* operation. The *NNI* distance from one tree to another is defined as the minimum number of *NNI* operations required to transform one tree into the other [115, 41].

1.3.2 Subtree Prune Regraft Metric

The Subtree Prune Regraft operation is a type of tree rearrangement operation which is useful for studying the changes of the shape of a tree due to recombination [7]. An *SPR*

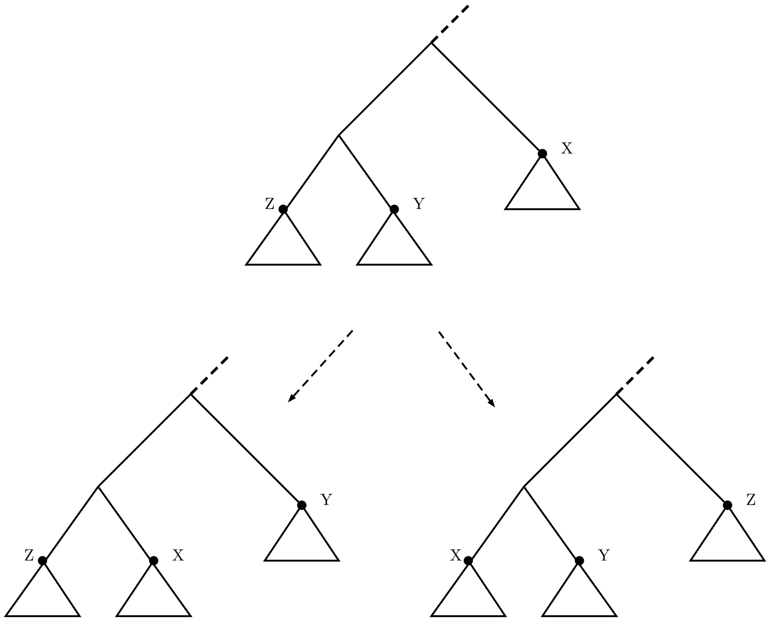


Figure 1.3: Two possible configurations for an NNI move on a rooted tree.

move on a labeled phylogenetic tree T involves cutting an edge from a tree T, which results in pruning a subtree t from T, and then "regrafting" t onto an arbitrary location in the remaining part of T [178]. The *SPR* distance between two labeled phylogenetic trees is the minimal number of *SPR* moves needed to change one tree into another.

1.3.3 Robinson-Foulds Metric

Given an unrooted labeled phylogenetic tree T with n tips, consider an edge $e \in E$. Removing e from the tree induces a bipartition on the set of the tips of the tree $V = \{1, 2, ..., n\}$, each part of which corresponds to the labels of the tips of the two connected components obtained after removing e. Let us denote by $c(T)$ the set of all bipartitions producible by removing an edge of T. Since in an unrooted binary tree with n tips there are $(2n-3)$ edges, we conclude that $|c(T)| = 2n - 3$.

Consider two unrooted labeled phylogenetic trees T_1 and T_2. The *RF* distance is defined as the size of the *symmetric difference* between their bipartitions:

$$d_{RF}(T_1, T_2) = |c(T_1) \bigtriangleup c(T_2)| = |c(T_1)| + |c(T_2)| - 2|c(T_1) \cap c(T_2)|.$$

The *RF* distance between two unrooted labeled trees is always an even number since $|c(T_1)| = |c(T_2)| = 2n - 3$. The maximum value of *RF* distance is given by the number of internal edges of the trees. Since the number of internal edges in an unrooted labeled phylogenetic tree is $(n-3)$, the maximum value of *RF* distance is $2(n-3)$. One can also define a *normalized RF* distance by dividing d_{RF} by its maximum value, $2(n-3)$. However, we focus on the regular (unnormalized) RF distance in this thesis.

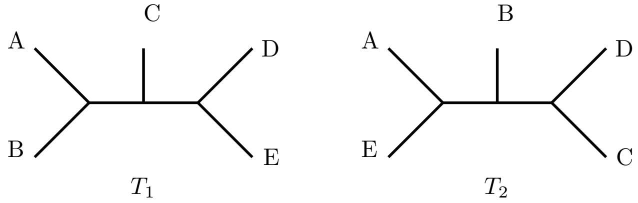


Figure 1.4: This figure shows two binary unrooted phylogenetic trees with 5 tips. The Robinson-Foulds distance between these two trees is 4.

1.3.4 Quartet Metric

The quartet metric compares two unrooted labeled trees based on the configurations of the quartets of nodes in each tree. The possible quartet configuration of four species in an unrooted labeled phylogenetic tree are shown in Figure 1.5.

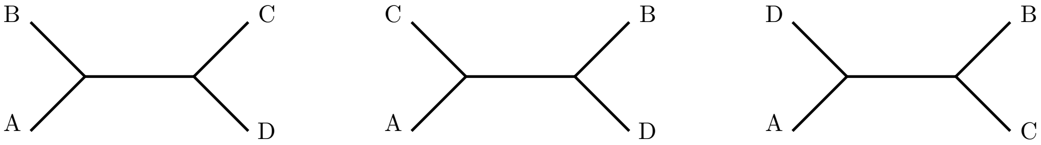


Figure 1.5: The three possible quartet topologies of species A, B, C, and D.

There are $\binom{n}{4}$ groups of four taxa in an unrooted labeled tree with n tips; for each of these groups, one of the three trees in Figure 1.5 will be consistent with a given tree. It is well-known that the complete set of quartet configurations is unique for a given tree and that the tree can be uniquely recovered from its set of quartet configurations in polynomial time [28]. This time is reduced to $O(n \log(n))$ for binary trees [110]. Given two evolutionary trees on the same set of n species, the quartet distance between them is the number of sets of four species for which the quartet topologies differ in the two trees. Brodal et al. [23] proposed an algorithm for computing the quartet distance between two binary trees in $O(n \log^2(n))$ time, which is the fastest algorithm for computing this metric so far.

1.4 Phylogenetic Reconstruction Methods

There are different methods to infer phylogenetic trees from multiple sequence alignments [192, 120, 166, 57, 64, 177, 53]. These methods are generally classified in two groups: character-based methods and distance-based methods. The first methods directly use the individual columns of aligned nucleotides or amino acids, while the other methods use measures of the overall differences between all pairs of sequences in the alignment (represented as a matrix of pairwise genetic distances). Each method has its own advantages and disadvantages, and the choice of a particular method depends primarily on a trade-off between the accuracy of the reconstruction and the time complexity of the method [44].

1.4.1 Distance-Based Methods

The most common algorithms in this group are: the unweighted pair grouping with arithmetic mean (*UPGMA*) [120], least squares (*LS*, also called minimum evolution) [57], neighbor-joining (*NJ*) [166]. The primary advantage of such distance-based methods is their computational speed, so they can handle large numbers of sequences (they also only require $O(n^2)$ space to store all pairs of distances, which n is the number of sequences). These methods are more practical in initial analysis of evolutionary relationships between sequences in a dataset. They are model-based, so assumptions are clearer. In contrast to their utility, one of the disadvantages of distance-based methods is to lose much of the potentially evolutionary informative information within an alignment by compressing it into sets of pairwise evolutionary distances [54, 44].

1.4.2 Character-Based Methods

One of the advantages of character-based methods over the distance-based methods is that the latter methods use all the information available in sequences at each homologous site. The most common methods in this group are the maximum parsimony (*MP*) [177], maximum likelihood (*ML*) methods [53], and Bayesian methods [78]. The main difference between likelihood-based methods (maximum likelihood methods or Bayesian methods) versus distance-based methods is that likelihood-based methods can estimate the molecular evolutionary model, whereas distance-based methods have to just assume it. Distance-based methods could use a highly complex model but would not be able to estimate the evolutionary parameters such as transition and transversion rates, rate heterogeneity across sites, strict or relaxed molecular clock and so on.

The Maximum Parsimony Method

In this method, the phylogeny of a group of species is inferred to be the branching pattern requiring the smallest number of evolutionary changes. A major issue regarding the maximum parsimony method is that there will usually be multiple equally parsimonious trees. This model also shows a statistical inconsistency in which it produces long branches in comparison with nearby branches. Maximum parsimony method generally implicitly considers a very simple model of evolution in which all possible nucleotide substitutions are equally probable. In this method, there is no need to search for branch lengths, so it is reasonably fast in comparison with other character-based methods. The result of maximum parsimony is reliable if the data is well structured and the number of nucleotide changes is small since this method can not account for multiple changes on the same branch. [177, 86].

The Maximum Likelihood Method

Maximum likelihood is a more complicated character-based method that reconstructs a phylogenetic tree that has the highest likelihood of being the correct representation of the phylogenetic relationships among the sequences. The primary advantage of the maximum likelihood method is that it is able to apply a wide variety of explicit evolutionary models. In particular different sites can have different substitution rates (which is also possible in distance-based methods), and these rates and other aspects of the molecular evolutionary model can be estimated along with the tree. The maximum likelihood method also tends to be robust to evolutionary model violation [188]. Another advantage of this method is that it evaluates different tree topologies. This method can be very computationally expensive for large datasets since it needs to enumerate of all the tree topologies. It is usually necessary to use approximate tree searching methods that find trees with high likelihoods, but will not always find the tree with the maximum likelihood. Another disadvantage of this method is that the result is dependent on the model of evolution that is used [53, 203]. Various software tools are available for phylogeny reconstruction using maximum likelihood which include RAxML [181], FastTree [148], IQtree [135], phyml [70].

The Bayesian Method

The Bayesian method is a statistical inference methodology to produce the most likely phylogenetic tree for a given data. It produces a sample from a distribution over all possible phylogenetic trees, where the probability of any given tree is given by the prior times the likelihood under the specified model. The Bayesian approach is a widely used method for tree reconstruction due to the advances in computing power, the integration of Markov chain Monte Carlo (*MCMC*) algorithms, and the availability of user-friendly software implementing sophisticated models of evolution. The advantages of the Bayesian approach over the traditional methods are that it gives the probability of the result, is able to incorporate complex models of evolution and can handle sources of uncertainty. The choice of the substitution model and the priors are the challenging parts in this method [78, 80].

“Finding a correct and accurate phylogenetic tree is generally an extremely difficult task” [143] and briefly there are the following sources of uncertainty in a phylogeny reconstruction:

- Choosing the right molecular markers. Both nucleotide and protein sequences can be used for reconstructing a phylogenetic tree, but the reconstructed tree from each of the options could be quite different. It is common to use nucleotide sequences for inferring the tree from very closely related organisms since they tend to evolve more rapidly than proteins. On the other hand, for more diverse groups of organisms, protein sequences are used for reconstructing the tree.

- Performing multiple sequence alignment. Multiple sequence alignment is used to identify regions of similarity that may indicate functional, structural or evolutionary relationships between a set of sequences. The aligned positions are assumed to be genealogically related. A low-quality alignment could introduce additional variation in phylogeny reconstruction (causing long branch lengths and muddying the relatedness signal). It also could miss true variation (causing short branch lengths and missing informative data that could resolve branches in the tree).

- Choosing a model of evolution. Evolutionary models are mathematical models used to describe the rates at which one nucleotide replaces another during evolution. A number of different evolutionary models have been proposed. Choosing the wrong model would give wrong likelihoods to proposed trees and can be misleading in representing the true evolution.

- Determining a tree reconstruction method. There are basically two types of phylogenetic methods, character-based methods and distance-based methods. Finding an appropriate method depends primarily on a trade-off between the accuracy of the reconstruction and the time complexity of the method.

1.5 Tree Shape Statistic

Phylogenetic trees are most often used in biology to study the historical relationship between a number of species or organisms. These trees contain both branch lengths and information in the form of the tree shape. In these trees, the leaves represent extant species, while the internal branches indicate hypothesized speciation events [183]. The shape of a phylogenetic tree reveals useful information about its growth process and can be used to infer the rates of species formation and extinction. Therefore, one of the main applications of phylogenetic trees is to study cladogenesis [149]. Measuring the degree of imbalance or asymmetry of a tree shape can provide support for the hypothesis that species have different potential for speciation [16]. Tree shape statistics are commonly used to quantify, with a single number, aspects of the phylogenetic relationships among a group of species or organisms. Primarily, tree shape statistics measure the degree of balance or imbalance of an unlabeled phylogenetic tree.

Three factors affect the degree of balance of an inferred tree. First, all evolutionary models include stochasticity, so the random fluctuations could produce balance patterns deviating from what is expected. Second, the accuracy of the method used to estimate the tree would affect balance. Third, the degree of balance or imbalance of a tree has some information about the macroevolutionary processes that produced it. In other words, variation among extant lineages in speciation and extinction rates results in phylogenetic trees that are less balanced than those produced from evolutionary models in which each extant species has the same probability of extinction or speciation [124].

Many different indices have been proposed in the literature so far to measure the degree of imbalance of a tree, and they differ in the calculation and to some minor extent in behavior [6, 17, 31, 93, 124, 145, 173, 185]. Among tree shape statistics, two of the most commonly used ones are the Sackin index [163, 173] and the Colless index [34]. The Sackin index is the average path length from a leaf to the root of the tree [16]. The Colless index is the sum of absolute values $|r - s|$ for all internal nodes, where r and s are the numbers of leaves on the left and right subtree of a node, respectively [16]. McKenzie and Steel [118] have proposed the use of the number of cherries, i.e. the number of nodes with two leaf descendants, as a tree shape statistic. We provide a list of the commonly used tree shape statistics in Table 5.1.

These indices have some properties in common, including that they summarize the shape of a tree in a single number and they ignore the branch length and just consider the distribution of tips across nodes [124]. The power of some imbalance statistics has been evaluated [75, 93, 115]. The studies concluded that the Sackin and Colless statistics are two of the most potent statistics in distinguishing between distributions on tree shapes.

Tree shape statistics have been used as tools to test stochastic models of evolution [124]. They also can be used in detecting mass extinction, adaptive radiations, measuring continuous variation in speciation, and extinction rates and associate changes in these rates with ecological and biological rates [124]. Tree shape statistics have also found applications in Phylodynamics, where recent research shows that tree structures are used to predict short-term growth and fitness and can help resolve disease transmission patterns [31, 103, 145, 60].

1.6 Organization of this Thesis

In chapter 2, we briefly review the literature around the four problems discussed in this thesis: evaluating the power of tree shape statistics, introducing new tree shape statistics, applications of tree shape statistics and methods for computing the distribution of Robinson-Foulds distance.

In chapter 3, we introduce a new resolution function based on the Laplacian matrix to evaluate the power of different tree shape statistics to distinguish between dissimilar trees. We show that the new resolution function requires less time and space in comparison with the previously proposed resolution function for tree shape statistics.

In chapter 4, we introduce two classes of tree shape statistics. The first one is the optimal linear combination of two existing statistics with respect to a resolution function. The other class of proposed statistics is inspired by network science. We propose tree shape summaries that are complementary to both asymmetry and the frequencies of small configurations using tools from network science, including diameter, average path length, and betweenness, closeness, and eigenvector centrality.

In chapter 5, we investigate the application of tree shape statistics in predicting the successful influenza strains that persist in the next year's influenza outbreak. We use longitudinally sampled phylogenetic trees based on hemagglutinin sequences from human influenza viruses, together with counts of epitope site polymorphisms in hemagglutinin, to predict which influenza virus strains are likely to be successful. We extract small groups of taxa (subtrees) and use a suite of features of these subtrees as key inputs to the machine learning tools. Using a range of training and testing strategies, including training on H3N2 and testing on H1N1, we find that successful prediction of future expansion of small subtrees is possible from these data.

In chapter 6, we modify the dynamic programming algorithm for computing the distribution of Robinson-Foulds distance for a given tree (the fastest known algorithm for computing this distribution) by leveraging the Number-Theoretic Transform (NTT). We improve the running time from $O(n^5)$ to $O(n^3 \log n)$, where n is the number of tips of the tree. In addition to its practical usefulness, our method represents a theoretical novelty, as it is, to our knowledge, one of the rare applications of the Number-Theoretic Transform for solving a computational biology problem.

Chapter 7 summarizes the introduced methods and results. We conclude the thesis by a discussion of our methods and findings and possible directions for future work.

Chapter 2

Literature Review

In this chapter, we review the existing methods related to each of the problems considered in this thesis: Evaluating the power of tree shape statistics, introducing new tree shape statistics, applications of tree shape statistics and finally computing the distribution of the Robinson-Foulds distance from a given tree.

2.1 Evaluating the Power of Tree Shape Statistics

The power of eight tree shape statistics: Sackin (S), δ_n^2, Colless (I), B_1, B_2, $\sum I'$, *Mean* I' , *Mean* I'_{10} in detecting nonrandom diversification was evaluated by Agapow et al. [4]. The first five of these statistics are studied well in the literature (see section 2.2), and the rest are introduced and discussed in [4, 62, 150, 152]. The imbalance of each internal node I' is defined as the ratio between the deviation of the bigger branch from the minimum value of its range, and the amplitude of that range:

$$I' = \frac{B - m}{M - m},$$

where m is the minimum value that B can take, and M is the maximum one [62]. Based on this definition three measures are introduced: $\sum I'$ which is the sum of I' over all nodes; *Mean* I' is the mean of I' over all nodes, and *Mean* I'_{10} is mean of I' over the 10 oldest nodes in a tree. In order to evaluate the power of these eight statistics, Agapow et al. [4] simulated two phylogenetic trees under two *non-ERM* models. In the first model, rates depend upon the value of an evolving trait, and in the second model a lineage's rate declines with time since the last speciation event it experienced. The distribution of these eight statistics under the *ERM* model was calculated for phylogenetic trees of sizes $8, 16, 32, 64$ and used as a reference to compare with the distribution of these statistics under age-dependent rates and trait-dependent rates. The result shows that the rank ordering of the different measures in terms of power varies with tree size and, more notably, with the process used to generate imbalance. The reason is that the two scenarios simulated by Agapow et al. [4] leave different imbalance signatures, and the various measures are most sensitive to imbalance in different parts of the tree. When rates are based on age, imbalance is spread quite uniformly between basal (nodes adjacent or close to the root) and distal nodes (nodes far away from the root) in the tree. When rates are based on trait values, however, imbalance is concentrated toward the root of the tree. Agapow et al. [4] showed that S and I are the most powerful tools to detect the deviation from the *ERM* model generally, but they were less powerful when applied to the age-dependent rates model. *Mean* I' has a reasonable power and it can be easily apply to non-binary trees. Another approach is suggested by Kirkpatrick et al. for testing the evolutionary hypothesis [93]. They suggested to consider two measures that are most sensitive to imbalance in different parts of a tree in conjunction with each other (S and I can be used in conjunction with either B_1 and $\sum I'$).

M. Blum et al. [16] evaluated the power of Sackin (S), D, and the frequency of subtrees ($f_n(z)$) (see section 2.2.2 for the definitions of D and $f_n(z)$) in rejecting the *ERM* model. The biased speciation model that they used assumed that the speciation rate of a lineage is equal to r, and when this node splits, one of its descendants is given the rate pr, and the other one is given $(1-p)r$ where p is fixed for the entire tree. They simulated this model for a different number of species $n = 30, 100, 200$ and different values of p. The result shows that $f_n(z)$ performs poorly, and this statistic is not recommended for hypothesis testing, while S and D perform well in rejecting the *ERM* model against the considered scenario [16].

Matsen [115] proposed a geometric approach to evaluate the power of a set of tree shape statistics: I, S, δ^2, B_1, B_2, I_2 (see section 2.2 for the definition of these indices). Their approach is based on the intuition that the value of a good statistic should be similar for similar trees, and different for trees with different shape. The geometric approach is completely different from other approaches for evaluating the power of tree shape statistics. The resolution function presented in [115] quantifies the ability of a statistic to differentiate between similar and different trees based on a given distance matrix, while the previous methods tested the discriminative power of statistics to distinguish among different macroevolutionary models. The resolution function introduced by Matsen [115] is defined as:

$$R_D(f) = \frac{-1}{2} x'_f D_s x_f \tag{2.1}$$

The vector x_f is the centered normalized vector of a statistic f for the set of n_T phylogenetic trees (the set of all binary rooted trees with n leaves), and D_s represents the component-wise matrix square of the distance matrix D. The geometric approach to evaluate the power of tree shape statistics is motivated by the statistical method of multidimensional scaling (*MDS*). The goal of *MDS* is to find a set of n points in k-dimensional Euclidean space in which the distance between each pair of objects with respect to a distance metric is well approximated by the Euclidean distance between the corresponding points. Let H denote the $n_T \times n_T$ “centering matrix”, defined by:

$$H := I - n_T^{-1} 1 1^t.$$

Here, 1 is a vector with every entry equal to one, and 1^t is the transpose of this vector. According to the Rayleigh quotient, $R_D(f)$ has an upper bound and lower bound which are the maximum and minimum eigenvalues of X_D, defined as:

$$X_D := \frac{-1}{2} H D_s H \tag{2.2}$$

The results show that the value of scaled resolution for Sackin and Colless is very close to one, which is the upper limit, and the scaled resolution of I_2 is substantially lower than the other statistics. These results are consistent with Agapow and Purvis's results which show that Sackin and Colless are the most powerful statistics [4].

2.2 Tree Shape Statistics

2.2.1 Classical Tree Shape Statistics

One of the most widely used tree shape statistics is Colless imbalance I (we use I_n to emphasize the value of I for a phylogenetic tree with n leaves) [34]. The Colless index is the sum of absolute values $|r - s|$ for all internal nodes, where r and s are the numbers of leaves in the left and right subtree of a node, respectively.

$$I = \sum_{i \in \mathcal{I}} |r_i - s_i|$$

Because the value of I is highly dependent on the size of a tree, usually the normalized version is used:

$$I = \frac{1}{\binom{n-1}{2}} \sum_{i \in \mathcal{I}} |r_i - s_i|$$

The Colless index gives more weight to the older nodes (nodes close to the root) because the value of $|r - s|$ is larger for those nodes [93]. It is possible to construct an alternative version of Colless which weights all nodes equally:

$$I_2 = \frac{\sum_{i \in \mathcal{I}, j > 2} \frac{|r_i - s_i|}{j-2}}{n-2}$$

where j is the number of tips subtended by each internal node [124].

Rogers [159] investigated the properties of Colless index. The minimum value of I is zero achieved only by completely balanced trees. The maximum value of I is $\sum_{i=1}^{n-2} i = \frac{(n-1)(n-2)}{2}$, which corresponds to the completely unbalanced tree (caterpillar) with n tips (see Figure 1.1

and section 1.1 for the definitions of the completely balanced and the completely unbalanced trees with n tips).

Using the recursions proposed by Slowinski [176] for computing probabilities of phylogenetic trees under the *ERM* and the *PDA* models, Rogers [159, 160] developed a set of recursion equations to compute the expected value, variance, skewness, and complete probability distribution of I under the these two models.

Assume that all tree shapes are ordered and numbered consecutively from 1 to n_T (the number of rooted unlabeled phylogenetic trees with n tips), so n_i corresponds to the i^{th} shape with n tips. The probability of the phylogenetic tree n_i under the *ERM* (equations 2.3 and 2.4) and the *PDA* (equations 2.5 and 2.6) can be computed recursively from the probabilities of two lower order shapes r_j and $(n-r)_k$ of which it is composed [176]. Under the *ERM* model:

$$P(n_i) = \frac{2}{n-1} P(r_j) P((n-r)_k) \tag{2.3}$$

if r_j and $(n-r)_k$ do not have the same shape, or

$$P(n_i) = \frac{1}{n-1} P(r_j)^2 \tag{2.4}$$

if r_j and $(n-r)_k$ are the same and $P(1_1) = 1$. Under the *PDA* model, the recursions are as follows:

$$P(n_i) = \binom{n}{r} \frac{T_r T_{n-r}}{T_n} P(r_j) P((n-r)_k) \tag{2.5}$$

if r_j and $(n-r)_k$ do not have the same shape, or

$$P(n_i) = \binom{n}{r} \frac{T_r^2}{2T_n} P(r_j)^2 \tag{2.6}$$

if r_j and $(n-r)_k$ are the same tree and $P(1_1) = 1$ and $T_n = \prod_{i=1}^{n-1}(2i-1)$ is the number of rooted labeled phylogenetic trees with n tips.

Assume that $P(n_i)$ is the probability of the i^{th} tree shape among all possible trees with n number of leaves, then $E(I_n^m)$, the m^{th} moments of I for a phylogenetic tree with n leaves, can be calculated from equation 2.7:

$$E(I_n^m) = \sum_{i=1}^{n_T} I_i^m P(n_i), \tag{2.7}$$

where n_T is the number of shapes with n leaves and m can be any positive integer [141]. The value of Colless index for the shape n_i (I_{n_i}) can be computed recursively from the values of the two major subtrees r_j and $(n-r)_k$ that form it:

$$I_{n_i} = I_{r_j} + I_{(n-r)_k} + |n-2r| \tag{2.8}$$

Combining equations 2.3 – 2.7 with equation 2.8 would result in a recursion equation for computing the moments of I_n under the *ERM* and the *PDA* (equations 2.9 and 2.10 respectively):

$$E(I_n^m) = \frac{1}{n-1}\sum_{r=1}^{n-1}\sum_{j=1}^{n_r}\sum_{k=1}^{n_{n-r}} P(r_j)P((n-r)_k)(I_{r_j} + I_{(n-r)_k} + |n-2r|^m) \tag{2.9}$$

$$E(I_n^m) = \frac{1}{2}\sum_{r=1}^{n-1}\sum_{j=1}^{n_r}\sum_{k=1}^{n_{n-r}} \binom{n}{r}\frac{T_rT_{n-r}}{T_n} P(r_j)P((n-r)_k)(I_{r_j} + I_{(n-r)_k} + |n-2r|^m) \tag{2.10}$$

The first moment of I_n is the expected value or mean of I_n. The variance and the skewness of I_n can be calculated from the first two moments and the first three moments of I_n respectively. The equations for computing the mean, variance, and skewness using the moments for I_n and two other statistics are discussed later in this section. The recursion equations for computing the first moment (expected value), second and third moments of I_n under the *ERM* model are defined in equations 2.11 – 2.13, and these recursions for computing the moments under the *PDA* model are defined in equations 2.14 – 2.16 respectively [159, 160].

Under the *ERM*:

$$E(I_n) = \bar{I_n} = \frac{1}{n-1}\sum_{r=1}^{n-1}(2\bar{I_r} + |n-2r|) \tag{2.11}$$

$$E(I_n^2) = \frac{1}{n-1}\sum_{r=1}^{n-1}(2E(I_r^2) + 2E(I_r)E(I_{n-r}) + 4|n-2r|E(I_r) + |n-2r|^2) \tag{2.12}$$

$$\begin{aligned} E(I_n^3) = \frac{1}{n-1}\sum_{r=1}^{n-1}(2E(I_r^3) + 6E(I_r^2)E(I_{n-r}) + 6|n-2r|(E(I_r^2) + \\ E(I_r)E(I_{n-r})) + 6|n-2r|^2E(I_r) + |n-2r|^3) \end{aligned} \tag{2.13}$$

Under the *PDE*:

$$E(I_n) = \bar{I_n} = \frac{n!}{2T_n}\sum_{r=1}^{n-1}\frac{T_rT_{n-r}}{r!(n-r)!}(2\bar{I_r} + |n-2r|) \tag{2.14}$$

$$E(I_n^2) = \frac{n!}{2T_n}\sum_{r=1}^{n-1}\frac{T_rT_{n-r}}{r!(n-r)!}(2E(I_r^2) + 2E(I_r)E(I_{n-r}) + 4|n-2r|E(I_r) + |n-2r|^2) \tag{2.15}$$

$$E(I_n^3) = \frac{n!}{2T_n} \sum_{r=1}^{n-1} \frac{T_r T_{n-r}}{r!(n-r)!} (2E(I_r^3) + 6E(I_r^2)E(I_{n-r}) + \tag{2.16}$$
$$6|n-2r|(E(I_r^2) + E(I_r)E(I_{n-r})) + 6|n-2r|^2 E(I_r) + |n-2r|^3)$$

The complete probability distribution of I is computed by combining the probabilities of the lower order I values. Any value of Colless index i for a tree with n tips can be computed recursively from the Colless values of its two major subtrees with r and $n-r$ tips [160]. Assuming the Colless values of these subtrees are j and k respectively, we have:

$$i = j + k + |r - (n-r)| = j + k + |n - 2r| \tag{2.17}$$

Let $P(i|n)$ denotes the probability of imbalance value i for a tree with n leaves. It can be computed by summing the products of all pairs of lower order imbalance values that satisfy equation 2.17 and then normalizing the sum such that $\sum P(i|n) = 1$. The probability distribution of I for the *ERM* and the *PDA* models are shown in equations 2.18 and 2.19 respectively:

$$P(i|n) = \frac{1}{n-1} \sum_{r=1}^{n-1} \sum_{j=0}^{a} P(j|r)P(k|n-r), \tag{2.18}$$

for $a = i - |n - 2r| \geq 0$.

$$P(i|n) = \frac{n!}{2T_n} \sum_{r=1}^{n-1} \frac{T_r T_{n-r}}{r!(n-r)!} \sum_{j=0}^{a} P(j|r)P(k|n-r), \tag{2.19}$$

for $a = i - |n - 2r| \geq 0$.

The probability distributions and the expected values of I under the null models give us references to compare the degree of balance of a given tree [160].

Rogers [160] computed the moments of I under the *ERM* and the *PDA* models for simulated trees of size $4 - 100$ and shows that the mean and the standard deviation of I decreases rapidly as the number of tips increases. The skewness is initially negative but becomes positive quickly under both models. These results also show that under both models, the shape of a phylogenetic tree tends to go towards symmetric trees rather than asymmetric ones as the number of tips increases.

Another well studied tree shape statistics is Sackin index S (we use S_n in some cases to emphasize on the size of the tree). The Sackin index is defined as follows:

$$S = \sum_{j=1}^{n} N_j,$$

where N_j is the number of internal nodes in the unique path from tip j to the root of the tree. Usually, the normalized version of Sackin ($\bar{S}$) is used:

$$\bar{S} = \frac{1}{n}\sum_{j=1}^{n} N_j$$

An equivalent equation of Sackin' index is by computing the number of leaves under each internal node:

$$S = \sum_{i=1}^{n-1} N_i,$$

where N_i is the number of leaves under the subtree rooted at internal node i. The values of the Sackin index for the set of trees with n tips vary from $O(n\log(n))$ to $O(n^2)$. The minimum value is obtained by the completely balanced tree, and the maximum value corresponds to the completely unbalanced tree. Kirkpatrick and Slatkin [93, 18] showed that the expected value of S for a phylogenetic tree with n leaves can be computed as follows:

$$\mu(S) = E(S) = 2n\sum_{i=2}^{n}\frac{1}{i}$$

The value of S for more asymmetric trees is larger than its expected value under a null model, while the lower values imply the converse.

The variance of $\bar{S}$ (δ_n^2) is also used as a tree shape statistic and is defined as follows [93]:

$$\delta_n^2 = \frac{1}{n}\sum_{j=1}^{n}(N_j - \bar{S})^2$$

The expectation of δ_n^2 is not known analytically, but its minimum value equals to zero in the completely symmetric tree with n tips, and it reaches its maximum value in the completely asymmetric tree with n tips.

Rogers [161] suggested using the number of unbalanced nodes U in a tree as a tree shape statistics:

$$U = \sum_{i=1}^{n-1}[1 - \delta(r_i, s_i)], \tag{2.20}$$

where $\delta(x, y) = 1$ if $x = y$ and is 0 otherwise. r_i and s_i are the numbers of tips of two subtrees arising from internal node i of the tree. U and I are similar in terms of considering the number of tips of subtrees arising from internal nodes, but these two indices capture different aspects of a tree shape. So as mentioned in [161] the joint distribution of I and U can be used to test *ERM* hypothesis.

The method of using a recursion equation to compute the mean, variance, skewness and complete probability distribution [160] is extended in [161] to compute the moments and the probability distribution of Sackin index and the number of unbalanced nodes on a tree

U under the *ERM* and the *PDA* models. Recursion methods for computing the moments and probability distribution of individual variables can be used for any tree shape statistics that can be written recursively. Any value i of I, U, and S for a phylogenetic tree with n tips can be computed recursively from the lower order values j and k for lower tree shapes with r and $n-r$ tips, respectively:

$$i = j + k + g(r), \tag{2.21}$$

where $g(r) = |n - 2r|$, $1 - \delta(r, n-r)$, n for I, U and S respectively. The probability distribution of S and U can be computed recursively using this recursion as mentioned for the Colless index:

$$P(i|n) = \sum_{r=1}^{n-1} w(r) \sum_{j=0}^{\infty} P(j|r)P(k|n-r), \tag{2.22}$$

where $P(i|n)$ represents the probability of imbalance value i for a phylogenetic tree with n tips, and $w(r)$ is a normalizing function to make $\sum P(i|n) = 1$ (note that equation 2.22 is the generalization of equation 2.18 and 2.19) [161].

Let Z represents I, S, and U for a phylogenetic tree with n tips. In order to compute the moments and the probability distribution of Z, Rogers [161] defined the moment generating function using the definition of probability distribution in equation 2.22.

$$F(x|n) = \sum_{a=0}^{\infty} P(a|n)e^{ax} \tag{2.23}$$

$F(x|n)$ can be computed recursively by using equations 2.22 and 2.23:

$$F(x|n) = \sum_{r=1}^{n-1} w(r)F(x|r)F(x|n-r)e^{g(r)x}, \; F(x|1) \equiv 1 \tag{2.24}$$

All of the moments of an imbalance index can be computed from the derivatives of this generating function at $x = 0$.

$$E(Z^m) = \frac{d^m}{dx^m} F(0|n) \tag{2.25}$$

The expectation or mean (μ), variance δ^2, and standardized skewness G_1 of Z can be computed as follows:

$$\mu(Z) = E(Z) \tag{2.26}$$

$$\delta^2(Z) = E(Z^2) - E(Z)^2 \tag{2.27}$$

$$G_1(Z) = \frac{E(Z^3) - 3E(Z^2)E(Z) + 2E(Z)^3}{\delta^2(Z)^{3/2}} \tag{2.28}$$

The moments of Z can be computed using equations $2.11 - 2.16$. The mean, variance, standard deviation, and skewness for I, S and, U are computed using these recursion equations under the *ERM* and the *PDA* models for simulated trees of $4 - 50$ tips [161]. Comparisons between the moments of these imbalance indexes show that Colless and Sackin are highly correlated but U behave differently. A recursion equation for computing the joint distribution and moments of any pair of imbalance coefficients is introduced in [161]. The results of investigating the joint distribution between the pairs of statistics (I, S), (S, U), and (I, U) for trees of size $4 - 50$ indicate that the correlations of I and U and S and U decrease as the number of tips increases under both the *ERM* and the *PDA* models. In contrast, the correlation of I and S increases continuously. The results of computing the joint distribution of I and U also show that this distribution is highly informative for assessing the degree of imbalance of a tree and discriminating among different non-*ERM* models of macroevolution [161].

The Sackin index can also be normalized using its expected value:

$$\tilde{S} = \frac{S - E[S]}{n}$$

The distribution of the normalized Sackin index $\tilde{S}$ and the number of subtrees of given sizes are investigated in [16]. M. Blum et al. used the one to one correspondence between binary search trees and the Yule trees [5] to find the precise description of the limiting distribution of the Sackin index and the number of subtrees under the *ERM* model. The limiting distribution of the Sackin index for a large n is non-Gaussian (in contrast with the result by Slatkin et al. [93]) and can be defined as the solution of a functional fixed-point equation. Computing the variance of the Sackin index is complicated, but it can be estimated by exploiting the fact that each Yule tree uniquely corresponds to a binary search tree. Using this fact, the Sackin statistic is equal to the number of comparisons used by quick sort algorithm to sort a random input [16]. A comparison between the expected value of the Sackin index and the average running time of quick sort algorithm is a simple proof that quick-sort takes $O(n \log(n))$ time on average.

The mean, variance, and covariance of the Sackin and Colless indices and their limiting joint distribution for large phylogenies under the *ERM* and *PDA* models have been computed asymptotically [16, 18, 93]. Blum et al. took advantage of the recursive structure of a phylogenetic tree and used the fixed-point method [82] for their analysis under the *ERM* model. Under the *PDA* model, the results are based on the connection between the uniform trees and Bernoulli excursions [187]. Under the *ERM* model, the expectation and the variance of Sackin and Colless index and the covariance of these two statistics for a phylogenetic tree with n leaves are shown in the following equations respectively [16, 18, 93]:

$$E(S_n) = 2n \log n + (2\gamma - 2)n + O(n) \tag{2.29}$$

$$Var(S_n) \sim (7 - 2\pi^2/3)n^2 \tag{2.30}$$

$$E(I_n) = n \log n + (\gamma - 1 - \log 2)n + O(n) \tag{2.31}$$

$$Var(I_n) \sim (3 - \pi^2/6 - \log 2)n^2 \tag{2.32}$$

$$Cor[S_n, I_n] \sim \frac{27 - 2\pi^2 - 6\log 2}{\sqrt{2(18 - \pi^2 - 6\log 2)(21 - 2\pi^2)}} \approx 0.98, \tag{2.33}$$

where $\gamma \approx 0.577$ is Euler's constant.

Under the *PDA* model of phylogenetic trees, the mean, variance, and covariance of the Sackin and the Colless indices are shown in the following equations respectively [18]:

$$E(S_n) \sim \sqrt{\pi} n^{3/2} \tag{2.34}$$

$$Var(S_n) \sim (10/3 - \pi)n^3 \tag{2.35}$$

$$E(I_n) \sim \sqrt{\pi} n^{3/2} \tag{2.36}$$

$$Var(I_n) \sim (\frac{10 - 3\pi}{3})n^3 \tag{2.37}$$

$$Cor[S_n, I_n] \sim 1 \tag{2.38}$$

The exact formula for the expected value of the Sackin index under the *PDA* model is also computed by Mitter et al. for the first time [121]. It was computed before for the limiting distribution of the Sackin index [18]:

$$E(S_n) = n(\frac{(2n-2)!!}{(2n-3)!!}) - 1$$

McKenzie and Steel [118] proposed a simple tree shape statistic: the number of cherries of a tree (a cherry is a pair of leaves that are adjacent to a common ancestor node). They studied the distribution of this statistic (Ch_n) under the *ERM* model [118], and showed that the mean and the variance of Ch_n are as follows:

$$E[Ch_n] = \frac{n}{3}$$

$$Var[Ch_n] = \frac{2n}{45}$$

The distribution of Ch_n is asymptotically normal as the number of leaves goes to infinity [16, 118].

$$\frac{Ch_n - n/3}{\sqrt{2n/45}} \to N(0,1)$$

Two measures B_1 and B_2 were introduced by Shao and Sokal [173]. B_1 explores the internal nodes of a tree, excluding the root and B_2 explores the external nodes. Let M_i represent the height of the subtree rooted at an internal node i and N_j represent the number of internal nodes on the unique path from tip j and the root of the tree, then B_1 and B_2 are defined as follows:

$$B_1 = \sum_{i \in \mathcal{I}} M_i^{-1}$$

$$B_2 = \sum_{j \in \mathcal{L}} \frac{N_j}{2^{N_j}}$$

Page [139] suggested another tree shape statistic R, which inspired by the scheme introduced by Furnas [61] for numbering the tree topologies: *"left-right rooted ranking"*. This scheme assigns a unique integer to each of the distinct topologies of trees with n tips, in which the completely unbalanced tree (caterpillar) has the smallest value, 1, and the most balanced tree would be assigned the largest possible value n_T (equal to the number of distinct tree topologies with n tips). The suggested statistic R is simply the position of the tree in the *"left-right rooted ranking"* order [139]. Kirkpatrick and Slatckin introduced a recursion to compute the value of R [93]:

$$R(T) = \begin{cases} \sum_{i=1}^{s-1} i_T(n-i)_T + [R(T_L) - 1]r_T + R(T_R) \qquad if \qquad r > s \\ \sum_{i=1}^{s-1} i_T(n-i)_T - s_T + R(T_L)[s_T - \frac{1}{2}R(T_L) + \frac{1}{2}] + R(T_R) \qquad if \qquad r = s, \end{cases}$$

where i_T is equal to the number of unlabeled rooted trees with i tips and can be computed recursively:

$$i_T = \sum_{j=1}^{\lfloor i/2 \rfloor} j_T(i-j)_T + \frac{1}{2}(\frac{i}{2})_T[(\frac{i}{2})_T + 1]$$

where $1_T = 2_T \equiv 1$ and $x_T = 0$ for non-integer x. T_R and T_L are the subtrees with r and s tips respectively and descending from the root.

2.2.2 Recently Proposed Tree Shape Statistics

Blum et al. [16] proposed a new statistic based on the comparison of the theoretical and empirical distributions of the number of subtrees of a tree under the *ERM* model. Let X_n denote the distribution of the number of leaves under a randomly chosen node in a phylogenetic tree T with n leaves under the *ERM* model and let $f_n(x)$ and $P_n(x)$ denote the frequency of subtrees of size x in T and the probability of $X_n = x$ respectively. The new statistic D is defined as follows:

$$D = \sum_{x=2}^{n} x|f_n(x) - P_n(x)|$$

M Blum et al. [16] estimated the quantiles of the distributions of this new statistic D using experimental design.

Matsen [116] introduced a practical genetic algorithm to optimize over a class of tree shape statistics called binary recursive tree shape statistics ($BRTSS$). The value of this group of statistics for a given tree can be computed recursively from its two main subtrees (the two subtrees emerged from the root). Let ρ define this recursion and λ be the base case:

$$s(T) = \begin{cases} \rho(s(X), s(Y)) & \text{if } T = XY \\ \lambda & \text{if } T \text{ is a leaf} \end{cases} \tag{2.39}$$

where XY represents X and Y as the two subtrees of T. For example, the number of leaves of a tree can be written as a recursive tree shape statistics with $\lambda = 0$ and $\rho(x, y) = x + y$. Generally, a BRTSS of length n on a tree is an ordered pair (λ, ρ) where $\lambda \in R^n$ and ρ is an n - vector of $Symm^2(R^n) \to R$. Here, $Symm^2(R^n)$ denotes the symmetric product of R^n with itself.

$$s_i(T) = \begin{cases} \rho_i(s_1(X), ..., s_n(X), s_1(Y), ..., s_n(Y)) & \text{if } T = XY \\ \lambda_i & \text{if } T \text{ is a leaf} \end{cases} \tag{2.40}$$

Using this definition Colless index can be written as a BRTSS of length 2 with the base cases $\lambda_1 = 0$ and $\lambda_2 = 1$ and the recursions:

$$\rho_1(x_1, x_2, y_1, y_2) = x_1 + y_1 + |x_2 - y_2|$$

$$\rho_2(x_1, x_2, y_1, y_2) = x_2 + y_2$$

The optimization method then will be applied on a set of BRTSS and modify λ and ρ through mutation and crossover to find the optimum statistics. A remarkable number of well known statistics are definable using expressions in the $BRTSS$. Matesn [116] showed the application of his proposed method by finding a new powerful statistic which works better than some of the classical statistics to distinguish between a set of trees from Aldous branching distribution [6] and a sample of trees from TreeBASE dataset (*ERM-solved* trees) [17]. The method proposed by Matsen [116] can be used to find a customized tree shape statistic for a certain application.

Arnau Mir et al. [121] introduced a new statistic called *total cophenetic index* Φ which is the sum of the cophenetic values of all pairs of different leaves (the cophenetic value of a pair of leaves i and j in a phylogenetic tree T is defined as the depth of their lowest common ancestor):

$$\Phi(T) = \sum_{1 \leq i < j \leq n} \phi_T(i, j)$$

This statistic has several advantages including that it can be used for an arbitrary tree (not just binary trees), can be computed in linear time and it has a broader range of values (up to $O(n^3)$ compare to $O(n^2)$ for Sackin and Colless imbalance). The higher resolution power of Φ in comparison with the Sackin and the Colless indices makes it a better candidate for evolutionary hypotheses testing. They[121] also computed its maximum and minimum values for arbitrary trees and binary trees as well as the exact formula for its expected value for binary trees under the *ERM* and *PDA* models. The maximum value of Φ for a set of trees with n leaves is obtained by the most unbalanced tree (caterpillar), and the minimum value for arbitrary trees and binary trees are reached at the star trees and the most balanced tree, respectively. The expected value of Φ under the *ERM* ($E_Y(\Phi)$) and the *PDA* ($E_U(\Phi)$) models are given in equations 2.41 and 2.42.

$$E_Y(\Phi) = n(n-1) - 2n(h_n - 1) = O(n^2) \tag{2.41}$$

Where h_n is the n^{th} harmonic number:

$$h_n = \sum_{i=1}^{n} \frac{1}{i}, n \geq 2$$

$$E_U(\Phi) \sim \frac{\sqrt{\pi}}{4} n^{5/2} \tag{2.42}$$

Another approach to study the imbalance of a tree was proposed by Aldous [6]. They represented a binary tree by a set of splits:

$$(m, i) = (\textit{size of parent clade}, \textit{size of the smaller daughter clade}),$$

which can be simply plotted as a scatter diagram. Given a phylogenetic tree, the proposed method is to estimate the median size of the smaller clade as a function of the size of the parent clade using some nonlinear regression. This function can be used to describe the degree of imbalance of a tree. Figure 2.1 is a log-log plot which illustrates the splits of a real dataset [6]. The plot also shows some lines giving the approximate median of the size of the smaller daughter clade predicted by the beta-splitting model for several values: $\beta = 0$ (Markov model), $\beta = -1.5$ (PDA model), $\beta = -1$, and $\beta = -\infty$. The model that best fits the data is the line in which about half of the points are above it, and the other half are below the line. Using this approach makes it possible to compare trees with a different number of leaves, which is not possible when simply using numerical summary statistics [6].

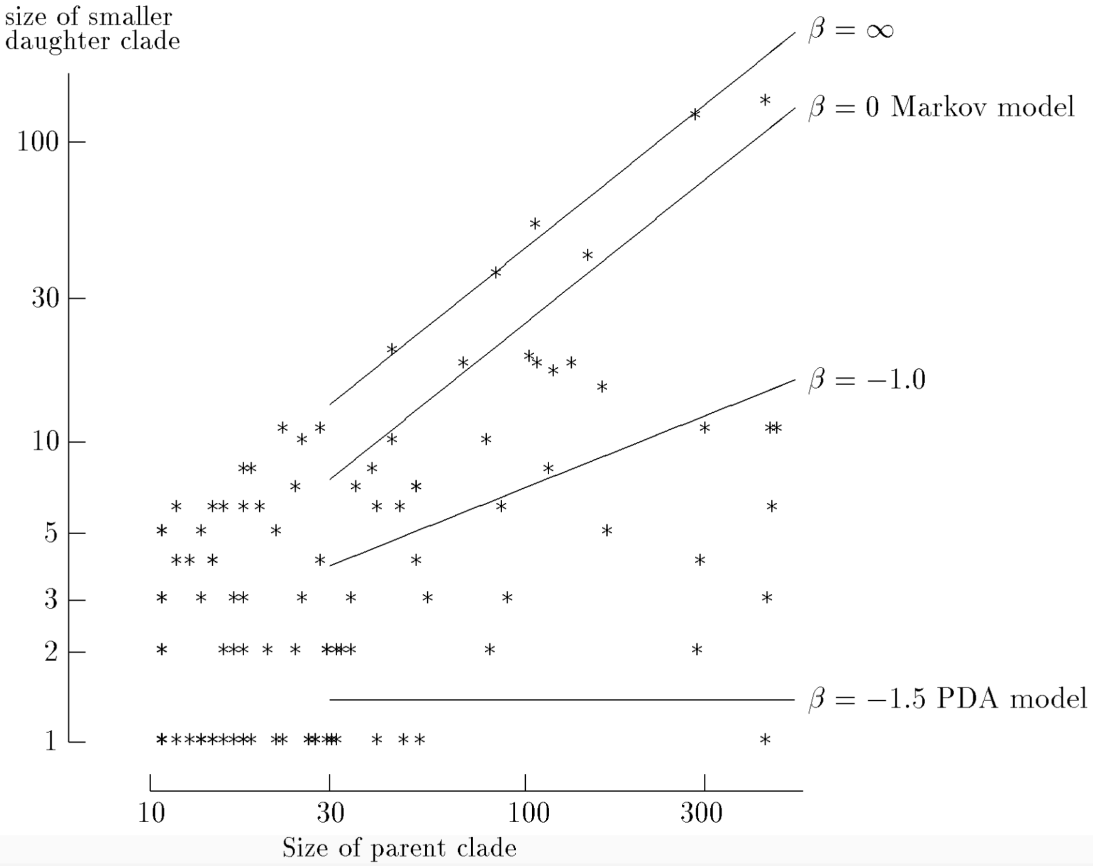


Figure 2.1: Log-log scale plot of the splits of a tree from a real dataset, and approximate median lines for the beta splitting model[6].

The list of commonly used tree shape statistics is provided in table 2.1.

Name	Description	Short form	Reference
Statistics inspired by network science			
Betweenness centrality	# of shortest paths through node	between	[30]
Closeness centrality	total distance to all other nodes	closeness	[30]
Eigenvector centrality	value in Perron-Frobenius vector	eigen	[30]
Diameter	largest distance between 2 nodes	diameter	[30]
WienerIndex	sum of all distances between 2 nodes	Wiener	[30]
Numbers of small configurations			
Cherry number	# of nodes with 2 tip children	cherries	[118]
Double cherries	# of nodes with 2 cherry children	doubcherries	[30]
Pitchforks	# of nodes with 3 tip descendants	pitchforks	[162]
Clades of size x	# of nodes with x tip descendants	num$x$	[162]
4-caterpillar	# of caterpillars with 4 tips	fourprong	[162]
Tree-wide summaries			
Colless imbalance	$\sum_{i\in\mathcal{I}} \lvert r_i - s_i \rvert$	colless	[34]
Sackin imbalance	mean path length from tip to root	sackin	[163]
Maximum height	max # of steps from the root	maxheight	[31]
Maximum width	max # of nodes at the same depth	maxwidth	[31]
Stairs1	the portion of imbalanced node	stairs1	[31]
Stairs2	the average of $\frac{min(r_i,s_i)}{max(r_i,s_i)}$ over the internal nodes.	stairs2	[31]
Max difference in widths	$\max_i(n_{i+1} - n_i)$	delW	[31]
Variance	variance of internal nodes depth	σ_n^2	
I_2	$\sum_{\substack{j\in\mathcal{I}\bigcup\{r\} \\ r_i+s_i>2}} \frac{\lvert r_i-s_i\rvert}{\lvert r_i+s_i-2\rvert}$	I_2	[115]
B1	$\sum_{i\in\mathcal{I}} M_i^{-1}$	B1	[115]
B2	$\sum_{i\in\mathcal{L}} \frac{N_i}{2^{N_i}}$	B2	[115]
ILnumber	# of internal nodes with a single tip child	ILnumber	[31]

Table 2.1: Brief definition for tree shape statistics. Here r_i and s_i are the number of tips of the left and right subtrees of an internal node respectively. n is the number of tips of a subtree. n_i is the number of nodes at depth i, M_i represents the height of the subtree rooted at an internal node i, and N_i is equal to the depth of tip i. A ladder in a tree is a set of consecutive nodes with one tip child. We represent the set of all internal nodes of a tree by $\mathcal{I}$, the set of all tips (or external nodes) by $\mathcal{L}$. The tree shape statistics induced by betweenness centrality, closeness centrality and eigenvector centrality are defined as the maximum values of each centrality over all the nodes of a tree, and distances are in units of number of edges (without branch lengths). We introduce the network statistics in chapter 4 of this thesis.

2.3 Applications of Tree Shape Statistics

We study the applications of tree shape statistics in two fields: the applications in testing the evolutionary hypotheses and the applications in phylodynamics.

2.3.1 Testing Evolutionary Hypotheses

Kirkpatrick and Slatkin [93] generated a large number of random trees with different numbers of tips and calculated the distribution of Sackin index, Colless index, σ_n^2, R, $B1$, and $B2$ for these trees under the *ERM* model. The first three of these statistics have larger values for more asymmetric trees. On the contrary, the last three ones have smaller values for more asymmetric trees. They then used these distributions to estimate the expected values of each statistic and their 95% confidence limits. Using this statistical method, one can identify a nonrandom pattern in a given phylogenetic tree of arbitrary size. Kirkpatrick and Slatkin [93] applied their method on two published phylogenies for groups of leaf beetles: Ophraella taken from Futuyma and McCafferty [63] and Phyllobrotica form Farrell and Mitter's [122]. Comparing the distribution of these six statistics inferred form these phylogenies against those under the random branching model shows that the Phyllobrotica phylogeny is significantly asymmetric by all of the six measures except B_1. The results also show that Ophraella phylogeny is significantly asymmetric by B_1 and is not significantly different from the random model by considering the other measures. These results show that different tree shape statistics measure different aspects of a tree shape, and a significant deviation from the confidence interval of one of the measures suggests a departure from random branching.

Blum et al. [17] used a tree shape statistics, F_n, to study the imbalance of a set of phylogenetic trees from the TreeBase dataset [199]:

$$F_n = \sum_{i=1}^{n-1} log(N_i - 1),$$

where N_i represents the number of leaves of the subtree rooted at node i. They showed that the statistical test based on F_n is the most powerful test for rejecting the *ERM* model against the *PDA* and conversely [17]. They performed maximum likelihood parameter estimation under beta-splitting model for three sets of trees: fully resolved, *ERM*-solved (solving polytomies by replacing non-binary nodes with *ERM*-like subtrees), *PDA*-solved (solving polytomies by replacing non-binary nodes with *PDA*-like subtrees). Their results showed that the *AB* model corresponds to $\beta = -1$ best fits the TreeBase data.

Some research has been conducted in order to design evolutionary models that can simulate trees with the same variation in diversification rate as real trees in the literature [124]. There are some reasons for the variation in speciation rates among lineages, including refractory periods after speciation, adaptive radiations, selective extinctions, and fluctuations in the environment causing differences in the selection pressures [185]. From a biological point of view, in balanced trees, all the clades have the same probability of speciating, while in unbalanced trees, species from a highly branching clade are more prone to speciate in the future. This memory of past success would be diminished due to the effect of the mutation during the time. Mutation would destroy the correlations between the properties of the distant species [185].

Stich et al. [185] analyzed the topological properties of phylogenetic trees generated by different models of evolving population (population of RNA sequences and a simple model with mutation and inheritance) and compared them with trees generated from a class of *ERM* model (Coalescent). Their approach was biologically motivated and differed from many models that have an algorithmic nature. In their model, the distance from each RNA secondary structure to a target secondary structure determines its ability to replicate. Two measures are used by Stich et al. [185] to quantify the shape of a tree: the subtree size and the cumulative branch size. For each node i in a tree the subtree size A_i is the number of nodes of the subtree rooted at i and the cumulative branch size C_i is defined: $C_i = \sum_j A_j$, where j runs over all nodes of the subtree rooted at i. For a given tree the probability distribution of A and C may demonstrate power law tails, $P(A) \sim A^{\alpha}$ and $P(C) \sim C^{\gamma}$. There is a one to one relationship between A and C of the scaling type: $C \sim A^{\mu}$, with $\mu = (1-\alpha)/(1-\gamma)$. The value of the exponents determines the degree of imbalance of a tree. For example, for a completely unbalanced tree, the value of α, γ and μ are 0, 1/2 and 2 respectively and for a completely balanced tree, these values are 2, 2 and 1 respectively. Their results show that the evolutionary parameters such as mutation rate or selection pressure do not have a significant effect on the scaling behavior of the trees, while the size of the trees profoundly affects the scaling exponents of the trees. They also showed that the shape of the trees generated from the biologically motivated models discussed in their work asymptotically agree with the completely balanced tree [185].

2.3.2 Applications in Phylodynamics

In addition to their evolutionary insights, recently, tree shape statistics have found application in Phylodaynamic. Phylodynamics is a new field which is at the intersection of phylogenetics and epidemic dynamics of viruses. Leventhal et al. [103] investigated the problem of whether the shape of a phylogenetic tree inferred from a pathogen population depends on the host contact structure underlying that tree. Three different contact structures are considered in their study. First, the *Erdös-Renyi (ER)* random graph [52]; in this model, individuals are connected with probability q. This model results in a graph with a Poisson degree distribution. Second, the *Barabási-Albert (BA)* graph [8]; in this model, each node is sequentially added to the graph and attached to k neighbors, where nodes that already have many neighbors have a higher probability of being connected to the new node. This model results in a degree distribution with a power law tail. Third, the *Watts-Strogatz (WS)* graph [198]; in this model, every node is connected to its k nearest neighbors. Each link is then updated with probability p in such a way that one end of the link is rewired to a randomly chosen node. Thus the degree of a node that loses the link decreases by one and the degree of a node that the link is rewired to increases by one. This process introduces shortcuts in the graph (i.e. decreases the mean shortest path). For $p = 0$ the graph has strongly connected communities. For $p = 1$ all links are randomly assigned and the graph is similar to the ER graph with the same mean number of neighbors (equal number of edges). For intermediate values of p, the graphs often display both strong community structure and short path lengths, which are characteristics of small-world graphs. Leventhal et al. [103] showed that simulations of epidemics on networks with non-random contact structure would result in transmission trees with topologies that exhibit significant differences from tree topologies that would be obtained under the assumption of random mixing. They also showed that quantitative measurements of tree shape such as the Sackin index can be used to differentiate between different classes of contact structures. The Sackin index can also be used to test whether the contact structure significantly deviates from what would be expected under random mixing.

Pompei et al. [145] investigated another application of the tree shape statistics in Phylodynamics. Different selective pressures of the host immune system cause the coexisting strains to have different fitness, which induces a different shape of reconstructed trees of RNA viruses [145]. They reconstructed and analyzed the phylogenetic trees of six RNA viruses: Human Flu H3N2 virus, the Avian Flu H5N1 virus, the Swine Flu H1N1, the Measles virus, the HIV-1 virus, both at the Intra-host and Inter-host level. Since tree shape statistics are highly dependent on the size of a tree, extracting information from the value of statistics of each tree can be misleading, and a ranking for the imbalance level of the six phylogenetic trees does not clearly emerge from the values of statistics. Pompei et al. [145] proposed a new methodology that deals with the dependency of tree shape statistics on the size of a tree. For each tree, they randomly select k independent sets of size n' and for all possible $2 \leq n' \leq n$ and extract the subtrees induced on these sets. For each set of subtrees with the same size they compute the value of three statistics. M: *The mean topological distance* [163]

$$M = \frac{1}{n} \sum_{j \in \mathcal{L}} N_j,$$

where the sum runs over the n leaves of the tree, and N_j is the number of nodes between the leaf and the root of the tree (note that M is a normalized Sackin index).
D: *The mean depth* [145]

$$D = \frac{1}{N} \sum_{i \in (\mathcal{L} \cup \mathcal{I})} N_i,$$

which is the mean topological distance of each node (internal nodes and leaves) from the root of the tree, and $N = 2n - 1$ (total number of internal nodes and leaves).

I": *The Asymmetry metrics* [145]

$$I" = \frac{1}{n-1} \sum_{j \in \mathcal{I}} (\max(r_j, s_j) - m_j)$$

This index inspired by the Colless index, and r_j and s_j defined the same as in the Colless index and m_j is the smallest integer not smaller than $(r_j + s_j)/2$. They then take the mean of the values of these three statistics for all subtrees of each tree. With this approach, they could investigate the dependence of the imbalance metrics on the size of the considered subtrees. The sampling of many subtrees with the same size allows for a better statistical analysis of their imbalance level and reduce the effect of noise and fluctuations. This procedure results in a plot with a different trend for different phylogenetic trees and gives us a clear ranking of the imbalance of each tree. They also extracted subtrees corresponded to a different time interval and computed the value of each statistic. Their results indicated that these statistics can differentiate the trees that have temporal properties like influenza trees [145].

Another problem in Phylodynamics is investigated by Frost et al. [60]; They consider how population structure affects the shape and the structure of a viral phylogeny in the absence of strong selection at the population level. They used the viral phylogeny of HIV-1 and developed a deterministic approximation to compute the number of tips, number of lineages, number of cherries and the Sackin index of the phylogeny over evolutionary time under two transmission models:

Simple model of HIV infection:

$$\frac{dS(t)}{dt} = \Lambda - \beta c S(t) \frac{I(t)}{N(t)} - \mu S(t)$$

$$\frac{dI(t)}{dt} = \beta c S(t) \frac{I(t)}{N(t)} - (\mu + \gamma) I(t),$$

where S denotes the number of susceptible individuals and I denotes the number of infected individuals, $N(t) = S(t) + I(t)$, β is the probability of infection per contact, c is the contact rate, μ is the natural mortality rate, γ represents the excess mortality caused by infection and Λ denotes the rate of immigration or birth of new susceptibles.

Models which include heterogeneity between individual:

The simple model of HIV infection described above, only considered one type of infected and susceptible individuals. There are more complicated models that include heterogeneity. Different kinds of heterogeneity include differences in infectivity at different times since infection, differences between hosts in contact rates, and geographical heterogeneity. Such heterogeneity can highly affect the transmission dynamics. Frost et al. [60] studied the behavior of transmission dynamics of the viral phylogeny of HIV-1 under two heterogeneous models in infectiousness and a model with risk structure. The first model is *Acute and chronic HIV infection* which infectiousness of HIV-1 is much higher in acute infection than chronic infection. The second model is *A model with risk structure* which they considered two groups of individuals with different contact rates and two assumptions for a fraction of contact between these two groups: the proportionate mixing and the preferred mixing.

Their results [60] show that the values of the considered statistics follow a different trend over time under these different models of HIV-1. Assortativity (a measure that captures how different types of individual cluster together on a tree) increased with higher infectiousness of acute infection. Sackin index showed the highest degree of imbalance for intermediate values of relative infectiousness of acute infection. In contrast, high sample fraction, asymmetry captured by the low number of cherries, was the greatest for high infectiousness during acute infection.

The host contact network structure has a significant influence on the dynamics of an outbreak. Colijn et al. [31] investigated the problem of whether the topological structures of phylogenetic trees contain information about the transmission patterns underlying an outbreak. Identifying the type of transmission patterns driving an outbreak can be very useful in controlling strategies and outbreak management. Three kinds of transmission patterns were investigated by Colijn et al. [31]: homogeneous transmission, transmission with a super-spreader, and chains of transmission. Different contact network patterns cause pathogen genomes to accrue mutations in different patterns, which results in observably different phylogenetic tree shapes. Colijn et al. [31] used a combination of a set of tree shape statistics to classify trees according to different transmission patterns underlying the trees. Their results demonstrate that phylogenetic tree structure can reveal transmission dynamics. They simulated disease transmission networks with three different underlying transmission patterns: homogeneous transmission, transmission with a super-spreader, and chains of transmission. Each simulation started with a single infectious host who infects a random number of secondary cases over his or her infectious period; each secondary case infects others, and so on until the desired maximum number of cases is reached. They then trained k-nearest-neighbor *(KNN)* classifiers to classify trees based on their transmission pattern. They also used a *10-fold* cross-validation *SVM* to resolve differences between homogeneous transmission versus super-spreader networks. They applied their method on data from two real-world outbreaks and their results in prediction of transmission patterns were consistent with known epidemiology [31].

2.4 Computing the Distribution of the Robinson-Foulds Distance

Hickey et al. and Steel et al. [76, 182] propose a method for computing the distribution of the *RF* distance between a given tree T and all the trees on the same number of tips and having the same labels, using generating functions. For a given unrooted binary phylogenetic tree T with n leaves, let $b_m(T)$ denotes the number of unrooted binary phylogenetic trees which are at a distance m from T. The generating function of $b_m(T)$ is defined as follows [76]:

$$\mathcal{B}(T, x) := \sum_{m \geq 0} b_m(T) x^m \tag{2.43}$$

A recursive formula for the generating function $\mathcal{B}(T, x)$ is given in [76]:

$$B(T, x) = xB(T/e, x) + (1 - x^2)B(T_1, x)B(T_2, x)$$

Here, e can be any internal edge, T/e represents the tree after contracting e, and T_1 and T_2 are the maximal subtrees of T with e as a pendant edge. The exponential number of subcases in the above recursion implies a non-polynomial time algorithm to compute the distribution of the RF distance [26, 76].

Bryant and Steel [26], whose work serves as the basis of our approach in chapter 6 of this thesis, have proposed a polynomial-time algorithm via a dynamic programming approach for computing the distribution of the RF distance from a given tree T [26]. They also showed that a Poisson distribution whose parameter depends on the number of cherries of T can approximate it well when n is large.

In chapter 6, we modify the dynamic programming algorithm proposed by Bryant and Steel [26] and improve the running time from $O(n^5)$ to $O(n^3 \log(n))$ by leveraging the Number-Theoretic Transform (NTT).

Chapter 3

Evaluating the Power of Tree Shape Statistics

Tree shape statistics are summary measures used to quantify some aspect of the shape of a phylogenetic tree. Several statistics have been proposed for measuring the level of asymmetry of an unlabeled rooted tree in the literature. Tree shape statistics differ in the way they are calculated, and, to some extent, in behaviour [4, 6, 17, 34, 93, 124, 145, 173, 185, 163, 173, 150, 152]. These statistics only depend on the shape of the tree, so leaf labels and branch lengths are ignored in their study. It is commonly believed that the evolutionary processes that have produced a phylogenetic tree are reflected in the shape of the tree [54]. Among imbalance-based statistics, two of the most commonly used ones are the Sackin index [163, 173] and the Colless index [34]. The Sackin index is the average path length from a leaf to the root of the tree [16]. The Colless index is the sum of absolute values $|r - s|$ for all internal nodes, where r and s are the number of leaves in the left and right subtree of a node, respectively [16].

The power of eight tree shape statistics, including the Sackin and the Colless indices, in detecting nonrandom diversification has been evaluated by Agapow et al. [4]. They simulated phylogenetic trees under two models. In the first model, evolution rates depended on the value of an evolving trait, and in the second model, a lineage's rate decreased as a function of the time since the last speciation event it experienced. The distributions of these eight statistics under the *ERM* model were calculated and used as a reference to compare with the distribution of these statistics under age-dependent rates and trait-dependent rates. The result shows that the rank ordering of the different measures in terms of discriminatory power varies with tree size and, more markedly, with the process used to generate imbalance. Indeed, the two scenarios simulated by Agapow et al. [4] leave different imbalance signatures, and different measures are more sensitive to imbalance in different parts of the tree. When the rates are based on age, the imbalance is spread fairly evenly between nodes close to the root and far away from the root. When the rates are based on trait values, however, the imbalance is concentrated around the root of the tree.

Blum et al. [16] evaluated the power of the Sackin index, the frequency of subtrees $f_n(z)$, and a statistic called D, based on the frequency of subtrees, in rejecting the ERM model. They used a biased speciation model with a fixed parameter p: given a lineage with speciation rate r that splits, one of the descendants gets the rate pr, and the other one $(1-p)r$. They simulated this model for a different number of species and different values of p. The result shows that $f_n(z)$ performs poorly, while the Sackin and D statistics are very powerful [16].

Matsen [115] proposed a geometric approach to evaluate the power of eight tree shape statistics: I_c, S_c, δ^2, B_1, B_2, I_2, A_1, and A_2. His approach was based on the intuition that the value of a good statistic should be similar for similar trees, and different for trees with different shape. His approach quantifies the ability of a statistic to differentiate between similar and different trees based on a given distance matrix. Our work builds on ideas from Matsen [115] for evaluating the power of a tree shape statistic.

The *resolution* is the operational definition of performance for tree shape statistics, and it measures the discriminatory power of a tree shape statistic. In this chapter, we propose a new resolution function based on the Laplacian matrix instead of the distance matrix. Since computing the Laplacian matrix is faster than computing the distance matrix of a graph, the overall time complexity is reduced in comparison with previous methods while producing comparable results. The lower time and space complexity of the new resolution function enables us to easily explore the space of trees with more leaves.

3.1 Problem Definition

Considering the wide range of tree shape statistics, there is a need to evaluate the discriminatory power of these different statistics in a systematic way. A geometric method for this purpose was introduced by Matsen [115], based on a matrix of pairwise distances between a set of trees with a given size. Here we are proposing a different approach, based on the closely related, but computationally more tractable Laplacian matrix. In this chapter, we only consider unlabeled rooted phylogenetic trees.

3.2 Methods

We evaluate the power of previously published statistics by using two different resolution functions, R_D and R_L, which we describe in this section.

3.2.1 Geometric Approach

A geometric resolution function has been proposed by Matsen for evaluating different tree shape statistics [115]. This resolution function is based on the intuition that the value of a good statistic should be similar for similar trees, and different for trees with different shape. This intuition is summarized in Figure 3.1. In this figure, two statistics are used to evaluate the space of trees with 9 tips. We embedded the set of rooted unlabeled trees with 9 tips on the two dimensional space using a multidimensional scaling (*MDS*) [98] and the nearest neighbor interchange (*NNI*) distances between them (we use a version of *NNI* distance that defined for unlabeled trees). We then colored the points according to the values of the Sackin and I_2 statistics. The clustering pattern induced by the values of the Sackin index in the top figure indicates that the Sackin index can distinguish between distant trees well. On the other hand, the bottom figure indicates that the values of the I_2 statistic are not necessarily different for different trees, and so it induces a random-looking color pattern on the set of trees with 9 tips.

3.2.2 NNI Metric for Unlabeled Trees

We used the *NNI* metric introduced in [115] which is defined for unlabeled trees. We use a specific numbering scheme [106] for numbering the unlabeled rooted phylogenetic trees to account for tree isomorphism (two trees are isomorphic if one of them can be obtained from other by a series of flips, i.e. by swapping left and right subtrees of a number of nodes). This scheme assigns a unique integer to each of the unlabeled rooted phylogenetic trees with n tips of which the completely unbalanced tree (caterpillar) has the smallest value, 1, and the most balanced tree would be assigned the largest possible value n_T (equal to the number of distinct unlabeled rooted tree with n tips). Unique integers are assigned to the canonical form of the tree shapes (the canonical form of an unlabeled rooted tree represents the tree in a way that the size of the left subtree is always less than or equal the size of the right subtree for all internal nodes and whenever the two subtrees are of the same size, the left subtree is always the lower-numbered one). A full description of the numbering scheme for a tree in canonical form is below.

We create a total order on the n_T trees on n tips by arranging them as follows:

1. If the left subtree of T_1 has fewer tips than the left subtree of T_2, then $T_1 < T_2$.

2. If the left subtree of T_1 has the same number of tips as the left subtree of T_2 but is not identical to it, then the comparison is determined by the comparison of the left subtrees in their corresponding orders.

3. If the left subtree of T_1 is the same as the left subtree of T_2, the comparison is determined by the comparison of the right subtrees in their corresponding orders.

The number of a tree T in canonical form is then simply the position of T in this total order.

In order to implement the unlabeled *NNI* metric, we used the *nni* function in the *R* statistical computing language [154]. For each given unlabeled rooted phylogenetic tree T with n tips, this function returns the set of all labeled rooted phylogenetic trees, S_L, which are a single *NNI* move apart from T. In order to extract the set of all unique unlabeled rooted phylogenetic trees, S, from the result of the *nni* function, we first compute the

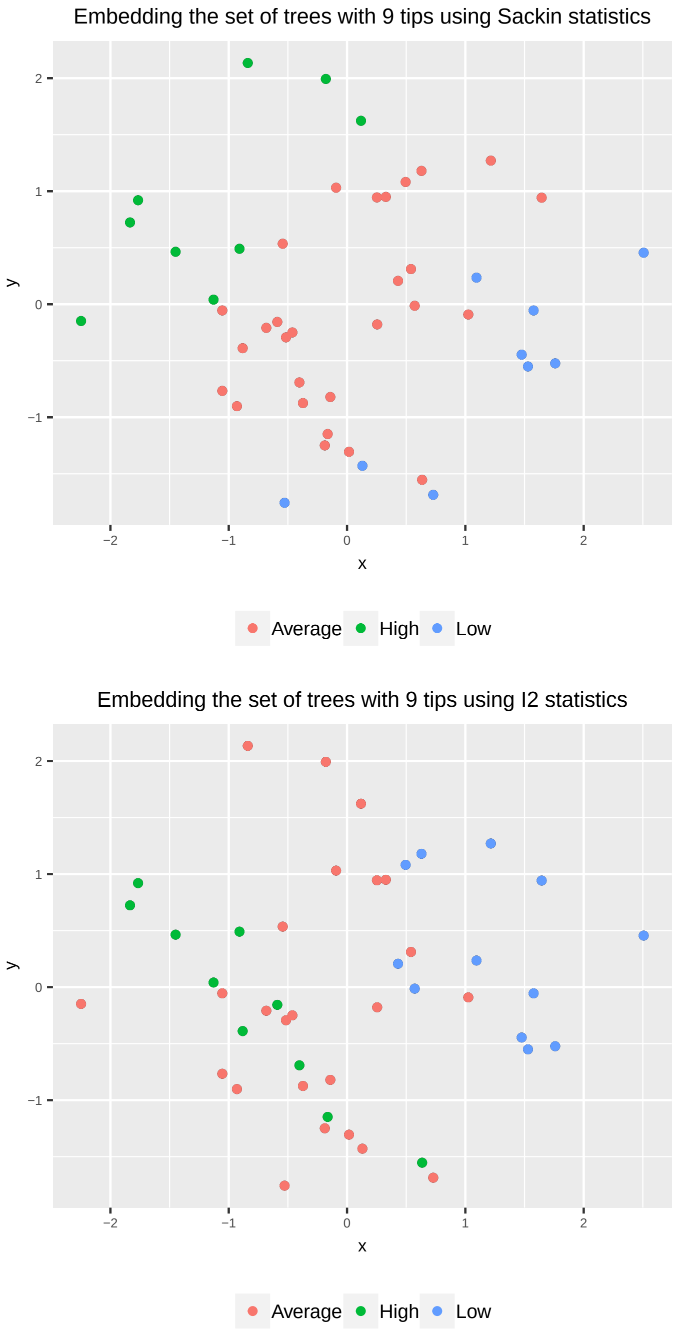


Figure 3.1: The geometric perspective of a good and bad tree shape statistic. From a geometric perspective, a good statistic can discriminate between different trees, and place similar trees together. In theses two figures, we embedded the set of trees with 9 tips using multidimensional scaling (*MDS*) and the *NNI* distance between the trees. The points in the top and bottom plots are colored based on their Sackin and I2 values respectively. The green, blue and red points correspond to the upper quartile, lower quartile, and the inter-quartile interval of the distribution of the statistics, respectively. The clustering pattern in the top figure indicates that the Sackin index can separate the trees into groups in a way consistent with the *NNI* distances, while the I2 index is unable to do so. This can also be seen in Table 3.1 using the resolution function for these two statistics.

canonical form of each tree in S_L. We then compute the labels of the canonical trees in S_L using the numbering scheme which results in a set of numbers. Finally we remove the duplicate numbers which results in a set of unique numbers representing the set of all unique unlabeled rooted phylogenetic trees that are at distance one from the given tree.

Note that the input phyolgenetic tree is unlabeled and the *nni* function assigns a random labeling to the tip set. We need to prove that a different labeling for the input tree would not affect the *NNI* distance between two unlabeled phylogenetic trees. Consider two phylogenetic trees with a single shape and two different sets of tip labels: T_1 and T_2. Assume that the tip labels of T_2 are a random permutation of the tip labels of T_1. Considering the definition of a single *nni* move (see Figure 1.3), the set of trees obtained from the *nni* function on T_2 is precisely the set of trees obtained from the *nni* function on T_1 but with the permutation applied to the tip labels. Since the numbering scheme ignores the tip labels and just considers the shape of a tree, applying the numbering scheme to both sets results in the same set of numbers (see Figure 3.2).

In the next step, we create the *NNI* graph such that each node is an unlabeled rooted phylogenetic tree with n tips and two trees are connected if they are in a distance one from each other (in the previous step, for each node, we found the set of all unique unlabeled rooted trees that are in distance one from the tree). The *NNI* distance between each pair of unlabeled trees with n tips is defined as the length of the shortest path between the corresponding nodes of the graph. We use Djikstra's well-known algorithm to calculate these shortest paths (Figure 3.3).

Lemma 1. *The NNI metric defined above for unlabeled trees satisfies the distance metric properties.*

Proof. A metric g is a non-negative and real-valued function on pairs of objects in a collection (called a metric space) M such that three constraints are met:

1. Identity: $g(x, y) = 0$ if and only if $x = y$

2. Symmetry: $g(x, y) = g(y, x)$ for all $x, y \in M$

3. Triangle inequality: $g(x, y) + g(y, z) \geq g(x, z)$ for all $x, y, z \in M$

The unlabeled *NNI* distance is a special case of the shortest-path metric on a graph and therefore it satisfies the above conditions. □

3.2.3 Resolution of Tree Shape Statistics Based on a Distance Matrix

Let n denote the number of leaves, let n_T denote the number of possible trees on n leaves, let d_{ij} denote the unlabeled *NNI* distance between trees i and j, and let H denote the $n_T \times n_T$ "centering matrix", defined by:

$$H := I - \frac{1}{n_T} 11^t.$$

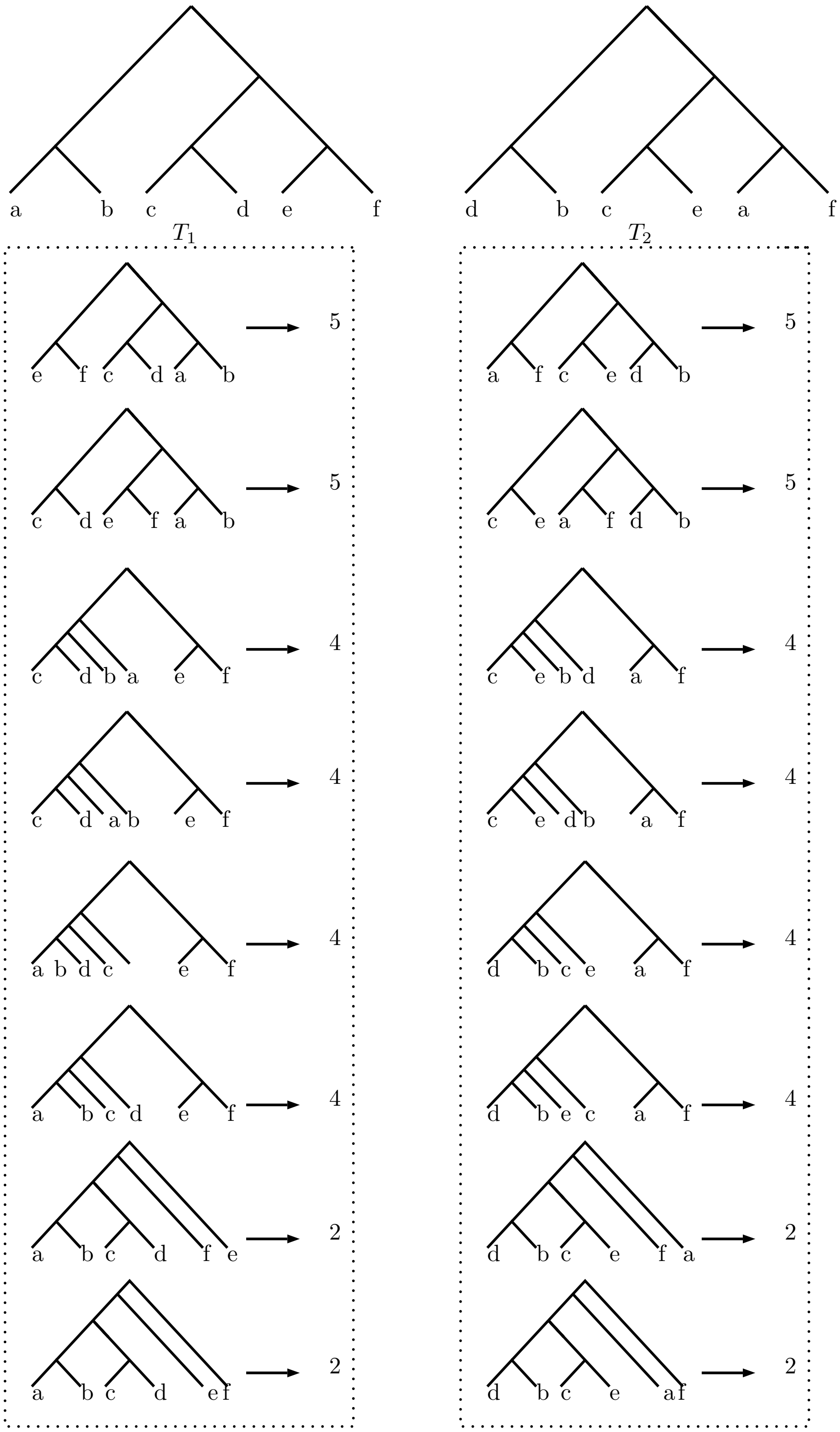


Figure 3.2: This figure shows that the initial random tip labels assigned by the *nni* function would not affect the *NNI* distance between two unlabeled trees. Consider a single tree shape with two sets of tip labels. Each column shows the set of all labeled rooted trees that are at distance one from the corresponding tree. Each tree shape is assigned a unique number using a numbering scheme. Since the numbering scheme ignores the tip labels, the assigned numbers in these two sets are exactly the same. The set of all unique unlabeled rooted trees which are at distance one from the given tree is computed by removing the duplicates from the set of numbers.

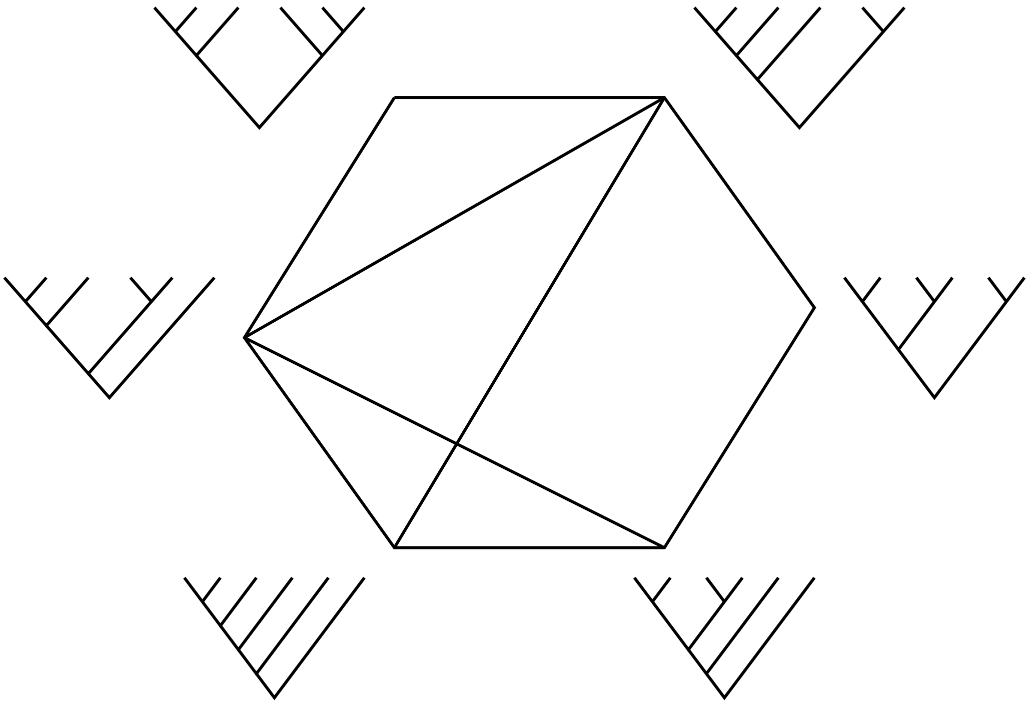


Figure 3.3: This figure depicts the definition of unlabeled *NNI* metric. Each node is a binary unlabeled rooted tree with 6 tips. Two trees are connected with an edge if they are one single *NNI* move apart. (Figure is taken from [115])

Here, 1 is a vector with every entry equal to one and 1^t is the transpose of this vector. The application of the centering matrix to a vector results in subtracting the mean from every component of the vector.

Assume that we are given a tree shape statistic f, and let y_f be a vector of size n_T whose i^{th} component is the value of the statistic f for the i^{th} tree. Assume that f is not constant on all the trees so that we can define the centered normalized vector of statistics x_f for the n_T trees as follows:

$$x_f := Hy_f/\|Hy_f\|$$

The resolution of the statistic f with respect to a distance matrix $D = (d_{ij})$ (any metric defined on the set of unlabeled rooted trees such as *NNI* or *SPR* distances) is defined in equation (3.1) [115]:

$$R_D(f) := \frac{1}{2}\sum_{i,j} -{d_{ij}}^2(x_f)_i(x_f)_j = -\frac{1}{2}x_f^t D_s x_f \tag{3.1}$$

Here D_s represents the component-wise matrix square of D, so that the ij-th component of D_s is ${d_{ij}}^2$. The higher the resolution value of a statistic, the more powerful it is from the geometric perspective. The goal is to maximize $R_D(f)$. It is easy to see that each term

$-{d_{ij}}^2(x_f)_i(x_f)_j$ is maximized when x_{f_i} is very negative and x_{f_j} is very positive, or vice versa, which means the value of a good statistic is similar for similar trees and different for different trees. This summation is also weighted by the distance, which means that pairs of trees that are a large distance apart contribute more than pairs of trees that are a small distance apart [115].

The geometric resolution is motivated by the statistical method of multidimensional scaling (*MDS*) [98]. The *MDS* method looks for a set of points $p_1, ..., p_{n_T}$ in K-dimensional Euclidean space that minimize the discrepancy between the true distances and the Euclidean distances:

$$\left[\sum_{i<j}(d_{ij} - |p_i - p_j|)^2\right]^{1/2}$$

where d_{ij} is the distance between tree i and tree j in the given metric. The Euclidean distance between this set of points approximates the distance between the trees. To find the optimal points in K-dimensional Euclidean space, the eigenvectors corresponding to the top K eigenvalues of $X_D = -\frac{1}{2}HD_sH$ are used [98, 115].

3.2.4 Resolution of Tree Shape Statistics Based on the Laplacian Matrix

In this section, we propose a new resolution function based on the Laplacian matrix instead of the distance matrix. Since computing the Laplacian matrix is faster than computing the distance matrix of a graph, the overall time complexity is reduced compared to the previous method.

The Laplacian matrix (L) is a matrix representation of a graph and is defined as follows:

$$L(i,j) = \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

For a given statistics vector f and the Laplacian matrix L of the graph on all trees with edges between trees at a distance of 1, we define our new resolution function in equation (3.2):

$$R_L(f) = x_f^t L x_f \tag{3.2}$$

Analogously to the previous section, x_f is the centered normalized vector of the given statistic vector y_f.

In contrast with the previous resolution function, for which a higher resolution value indicates a better statistic, here a good statistic has a lower resolution value. As follows from the definition of $R_L(f)$, we consider only pairs of trees which are adjacent when computing it. Since adjacent trees have similar topologies, a good statistic should assign similar values to them, so the value of the resolution for this statistic should be small.

An alternative interpretation of the Laplacian resolution is based on the idea of energy minimization, inspired by the use of the Laplacian for graph embedding [69]. It follows from the definition of L that, for any vector x,

$$R_L(x) = \frac{x^t L x}{x^t x} = \frac{1}{x^t x} \sum_{i \sim j} (x_i - x_j)^2, \tag{3.3}$$

where $i \sim j$ means that i and j are neighbors in the graph. If we think of each tree i as being located on the real line according to the value x_i of its statistic, and of each pair of neighboring trees as being connected by an elastic spring with unit spring constant, the total energy of this spring is given by the resolution's numerator.

Noting that $x^t L x$ does not change if x is replaced by $x + c$ for any constant c, we can assume that x is a vector with mean 0. If x is such a vector, we also have $x^t x = nE[x^2] = E[x^2] - E[x]^2 = n\text{Var}(x)$, where Var denotes the variance. Furthermore, in this case, $\sum_{i,j}(x_i - x_j)^2 = 2\sum_i x_i^2 - 2\sum_i x_i \sum_j x_j = 2x^t x$. Thus, the Laplacian resolution of a statistic measures, up to a scalar factor, the fraction of the total energy of the statistic (or variance, if the statistic is transformed to have mean 0) that gets allocated to neighboring trees. A statistic that places similar trees nearby will have low energy (and low variance), and hence, a low resolution value.

3.2.5 The Upper and Lower Bounds

We now need to transform the resolution function in order to ensure it is always in the interval $[0, 1]$, for comparison purposes. Following Matsen's original work [115], we use the Rayleigh quotient to compute the extreme values of the resolution. For a given symmetric matrix M and nonzero vector x, the Rayleigh quotient $R(M, x)$ is defined as:

$$R(M, x) = \frac{x^t M x}{x^t x} \tag{3.4}$$

It follows from the Courant-Fischer theorem [67] that the minimum and the maximum values of Rayleigh quotient are equal to the smallest and largest eigenvalues of M, respectively.

It follows that $R_D(f)$ has an upper bound and lower bound which are the maximum and minimum eigenvalues of X_D, defined as:

$$X_D := -\frac{1}{2} H D_s H \tag{3.5}$$

The upper bound for $R_L(f)$ is equal to the largest eigenvalue of L. We note that the smallest eigenvalue of L is zero and occurs only for constant vectors $x = x_1 1$ since the Cayley graph is a connected graph, according to equation 3.3; furthermore, by the symmetry of L, any other eigenvectors are orthogonal to the constant vector 1. Therefore, the lower bound is equal to the second smallest eigenvalue of L, also known as the Fiedler value of the Cayley graph [56].

Having determined the extreme values min and max for the resolution function, we transform the value of the resolution for all statistics to the $[0, 1]$ interval by using the linear transformation $x \to \frac{x-\min}{\max - \min}$; the resulting value is referred to as the scaled resolution.

3.3 Results

The superiority of our resolution over the previous one is the lower time and space complexity of its evaluation, so we can easily explore the space of the unlabeled rooted trees up to 25 leaves; this could only be done for the space of unlabeled rooted trees up to 17 leaves with the previous method [115]. It takes $O(n_T \log n_T)$ time and space to compute the Laplacian matrix of the Cayley graph of the set of n_T trees with n leaves for the *NNI* distance (since $n \in O(\log n_T)$ and the number of non-zero entries in each row of the Laplacian matrix is the degree of the corresponding vertex, which is the number of trees that are within distance 1 from the given tree). On the other hand, computing the distance matrix for the same set of trees takes $O({n_T}^2)$ time and space. Given the exponential growth in the set of trees with a fixed number of leaves [73], we are able to go further by decreasing the running time and space complexity (see Figure 3.4).

Computing the *NNI* metric is *NP*-hard [41], and we have only computed it for the space of trees with at most 17 leaves. To compute the *NNI* distance between each pair of trees on n leaves, we use the *nni* command of the *phangorn* package [169] in the R statistical computing language [154], which produces the list of all trees at *NNI* distance 1 from a specified tree. We then create the Cayley graph using the *igraph* package [38]. This Cayley graph has a vertex for every tree on n leaves, and an edge connecting any two trees at distance 1 (i.e. a single *NNI* move apart). Finally, we compute the *NNI* distance between every pair of trees on n leaves by using an all-pairs shortest paths algorithm on the Cayley graph [24, 20, 153, 29, 140].

n	I	S	σ_n^2	I_2	B_1	B_2
7	0.0984	0.0933	0.1082	0.1115	0.1179	0.0989
8	0.0808	0.0955	0.1110	0.0893	0.1164	0.0965
9	0.0507	0.0566	0.0662	0.0680	0.0797	0.0653
10	0.0327	0.0379	0.0471	0.0535	0.0629	0.0451
11	0.0222	0.0255	0.0326	0.0458	0.0511	0.0348
12	0.0183	0.0217	0.0282	0.0429	0.0473	0.0304
13	0.0160	0.0185	0.0238	0.0413	0.0441	0.0283
14	0.0147	0.0170	0.0217	0.0400	0.0421	0.0265
15	0.0137	0.0157	0.0197	0.0390	0.0404	0.0256
16	0.0130	0.0148	0.0184	0.0380	0.0389	0.0246
17	0.0123	0.0140	0.0170	0.0370	0.0375	0.0238
18	0.0117	0.0132	0.0160	0.0358	0.0361	0.0229
19	0.0112	0.0126	0.0150	0.0349	0.0349	0.0222
20	0.0107	0.0120	0.0141	0.0339	0.0338	0.0216
21	0.0102	0.0114	0.0133	0.0329	0.0327	0.0209
22	0.0098	0.0109	0.0126	0.0319	0.0316	0.0203
23	0.0094	0.0105	0.0120	0.0311	0.0306	0.0197
24	0.0090	0.0100	0.0114	0.0302	0.0297	0.0192
25	0.0086	0.0096	0.0108	0.0294	0.0288	0.0186

Table 3.1: Scaled resolution scores for the classical tree shape statistics based on our resolution function. n is the number of leaves. The best classical statistic is the Colless index and the worst ones are B_1 and I_2 (the same ranking as for the previous resolution function). The highlighted values correspond to the best statistics in each row.

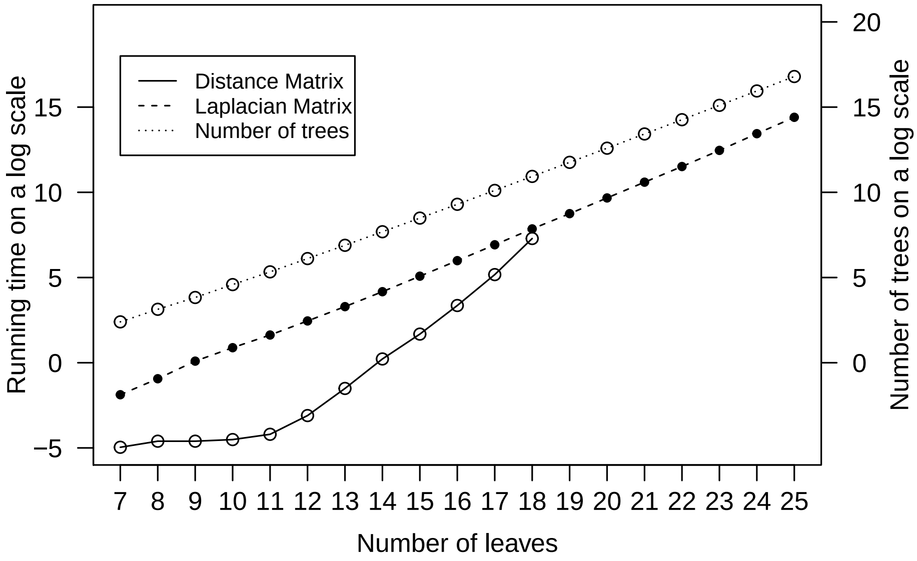


Figure 3.4: This plot shows the running time of computing the Laplacian matrix (dashed line) and the distance matrix (solid line) as well as the growth of the number of trees by increasing the number of leaves (dot line). To compute the geometric resolution one must compute the distance matrix; however, to compute our proposed resolution one needs only compute the Laplacian matrix. The left axis shows the time on a log scale. The bottom axis shows the number of leaves, and the right axis shows the number of trees with a specific number of leaves. The base unit for time is the second and a base-10 log scale is used for the left and right axis. A comparison between the slopes of the dashed line and the solid line in the plot shows that the running time of our proposed resolution is much faster that that of the previous method. The computational experiments were conducted on an Intel Core i7 with 2.4 GHz 64-bit processor, 16.0 GB of RAM and macOS Sierra 64-bit as the Operating System.

Chapter 4

New Tree Shape Statistics

Tree shape statistics are commonly used to quantify, with a single number, aspects of the phylogenetic relationships among a group of species or organisms. There has been considerable interest over the years in comparing the shapes of phylogenetic trees in order to understand evolutionary processes [180, 176, 71, 151, 93, 124, 17, 201]. A tree's shape specifies its connectivity structure. The lengths of its branches typically reflect either the time or genetic distance between branching events. Following the observation that reconstructed evolutionary trees are more asymmetric than random models predict [5], there have been efforts to summarise tree asymmetry in trees reconstructed from data and relate it to predicted asymmetry in evolutionary and ecological models [93, 62, 6, 145, 185, 124, 17, 151, 113]. There is also interest in establishing whether taxa from two phylogenies might correspond to each other, for example in the context of parasites and hosts or fossils of different origins [66], and in comparing simulated trees with trees from data in epidemiology, for example using Approximate Bayesian Computation [146, 167]. These applications require quantitative tools to compare phylogenetic trees with different taxa, and they require summary features that are informative of the evolution or epidemiology being studied. .

Summaries of tree shape have often focused on either asymmetry, or the frequency of various configurations such as cherries or ladders [162]. Well-known measures of asymmetry include the Colless and Sackin imbalance [35, 163]. Asymmetry measures tend to be correlated with each other, and do not fully capture the shape of a tree [66, 115], leading to an interest in exploring other statistics, comparisons tools and metrics for this task [79, 147, 105, 66, 33, 167]. Some metric approaches directly find a distance (or similarity) measure between unlabelled phylogenies; others seek an optimal labelling for two unlabelled trees and use metrics for labelled trees, but this is not feasible for large trees. Metric approaches also do not lend themselves to interpretable descriptions of trees that can easily be connected with generative models of evolution or epidemiology.

In one particularly important application of tree shape statistics, namely, Approximate Bayesian Computation, one seeks simulation parameters that produce trees that are similar enough to data-derived trees [39]. This leaves a need for new statistics to quickly provide a good summary of tree shape. Matsen used optimization over binary recursive tree shape statistics to find statistics that distinguish between sets of trees [115]. This set the stage for broadening the set of quantities used to describe trees, and gives rise to natural questions like why a particular recursive statistic separates two groups, whether it only happens to separate those specific trees or acts as a good way to distinguish between trees that are meaningfully similar to those in the two groups. The binary recursive class may furthermore exclude features that capture global information about a tree, such as features derived from its eigenvalues when viewed as a graph, known as spectral features. As the size of datasets and range of applications increases, it is reasonable to assume that inference will be improved by expanding upon the tools available to summarize tree topologies [115, 167]. In applications like Approximate Bayesian Computation, as well as other related attempts at summarizing tree shapes, one does not necessarily seek to relate summaries to evolutionary mechanisms. Rather, the key objective is to distinguish between trees in different categories or scenarios in an efficient way by computing simple statistics.

Different tree shape statistics can capture different aspects of the structure of a tree. This implies that the linear combination of pairs of different statistics can provide more information about the structure of the tree and results in a more powerful statistic than each single statistic. We investigate the optimal linear combination of the classical tree shape statistics, including the Sackin index, the Colless index, B_1 and B_2 with respect to a geometric resolution function. We evaluate the power of our suggested statistics and show that they perform better than the traditional statistics in distinguishing between different trees.

Network science has become an important paradigm for describing structural (topological) features of networks and using them to understand complex systems, ranging from protein interactions to social systems [104, 134]. Network science is thus a potential source of many novel ways to characterize tree shape since a phylogeny can be interpreted as a simple type of network or graph – specifically, a connected acyclic undirected graph. Network science offers many network features that can be adapted to describe shape. Here, we tailor tools from network science to summarize phylogenetic tree topologies. We thereby develop tree shape summaries that are complementary to both asymmetry and the frequencies of small configurations. These new statistics are fast to compute and will scale well to describe the topologies of large trees. They can additionally be easily adapted to take branch lengths into account. In order to evaluate the power of our proposed network statistics, We apply these statistics, alongside some conventional tree statistics, to phylogenetic trees from three very different viruses (HIV, dengue fever and measles), from the same virus in different epidemiological scenarios (influenza A and HIV) and from simulation models known to produce trees with different shapes. Using supervised learning algorithms, we find that the statistics adapted from network science perform as well as or better than conventional statistics.

4.1 Problem Definition

In contrast to the wide application range of tree shape statistics, existing statistics often do not suffice to distinguish between important scenarios, such as trees corresponding to different viral pathogens or different geographical scales for the same pathogen. In this chapter, we seek to fill in this gap by proposing two classes of tree shape statistics. First, we introduce a new class of tree shape statistics, which are linear combinations of two existing statistics that are optimal with respect to a resolution function. Second, we use network science, a well-developed branch of data science, to inspire 5 novel classes of tree shape statistics: 2 tree-wide ones (diameter and mean path length between two nodes), and 3 node-specific ones (betweenness, closeness, and eigenvector centrality). In this chapter, we only consider unlabeled rooted phylogenetic trees

4.2 Methods

4.2.1 Linear Combination of the Classical Tree Shape Statistics

In this section we propose a new class of tree shape statistics which is a meta-statistic obtained by finding the linear combination of existing statistics that results in the optimal resolution.

Here we focus on the linear combination of the Sackin and Colless indices, which we call the Saless index: Saless $= \lambda S + I$.

We choose the value of λ to maximize the geometric resolution, and is different for trees with different numbers of leaves. Our experiments suggest (though we have not formally proven it) that λ will converge to a limiting value as the number of leaves goes to infinity.

The optimal value of λ is the $\text{argmax}_{\lambda}(R_D(Saless))$ defined as follows:

$$R_D(Saless) = \frac{(\lambda S + I)^t D_s (\lambda S + I)}{(\lambda S + I)^t (\lambda S + I)} \tag{4.1}$$

Here, $D = (d_{ij})$ is the *NNI* distance matrix introduced for unlabeled phylogenetic trees (see Section 3.2.2 for more details), and D_s represents the component-wise matrix square of D, so that the ij-th component of D_s is ${d_{ij}}^2$. If we call the numerator of the resolution f and the denominator g, the problem reduces to finding λ for which $\frac{f\prime}{g} = \frac{g\prime}{f}$. This condition simplifies to a quadratic equation, and by solving that equation we find the value of λ for trees with up to 17 leaves. In Appendix A, we show that the optimal value of λ is always real; it can sometimes be negative, though in the case of the *Saless* statistic (with respect to the geometric resolution) it always appears to be positive. These values are shown in Table 4.1.

Table 4.1 and Figure 4.1 suggest that the value of λ may converge to a limit as the number of leaves goes to infinity. However, we were unable to verify the plausibility of this

n	7	8	9	10	11	12	13	14	15	16	17
λ	5.77	0.11	2.38	0.92	1.07	1.21	1.43	1.26	1.3	1.27	1.32
R_D	0.931	0.926	0.923	0.943	0.955	0.956	0.957	0.957	0.957	0.956	0.956

Table 4.1: The value of λ and the resulting resolution R_D for trees with different number of leaves.

behavior by going beyond $n = 17$, as the number of trees, which is the size of the dense distance matrix D_s, grows exponentially with the number of leaves.

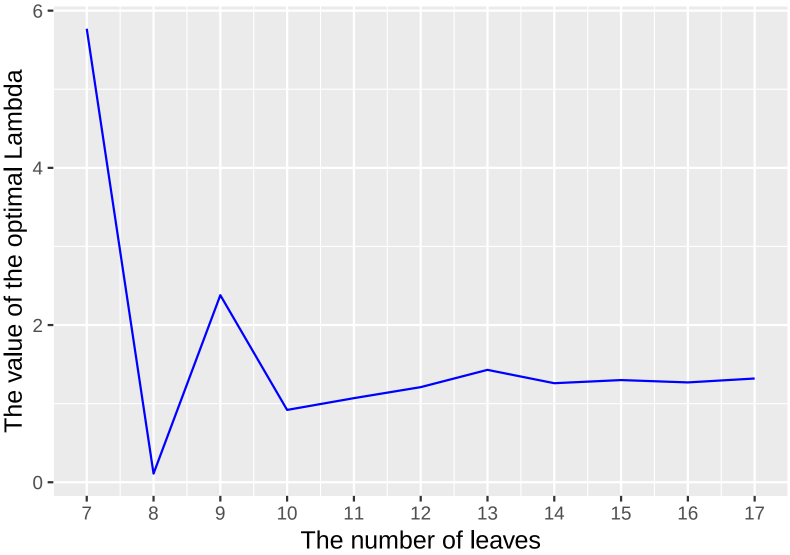


Figure 4.1: The value of λ appears to converge as the number of leaves grows.

In another experiment, we evaluated the combination of different pairs of statistics based on our new resolution function. We can show that the optimal coefficient λ gives the linear combination of the two statistics a better resolution than each individual statistic. This linear combination does not always result in an interpretable statistic, since the optimal λ is negative for some combinations. We also note that linear combination of statistics perform differently with different resolution functions. The results of these experiments are shown in Table 4.3.

4.2.2 Tree shape Summaries Based on Network Science

Network science, broadly defined as the study of complex networks, has produced a number of tools that have been applied in a variety of contexts [104, 134], including degree sequence, degree assortativity, density, diameter, and a number of node centrality measures. Only some of these apply naturally and informatively to phylogenies. Here we discuss those network science-inspired features that are informative for phylogenetic trees.

Diameter, Average Shortest Path and Wiener Index

The diameter (the maximum length of a shortest path between nodes) is a useful summary statistic. For general trees with N nodes (total internal and external nodes), it can be calculated in linear time using a “folklore” dynamic programming algorithm, and its value can vary between 2 for the star and $N-1$ for the path of length N. For phylogenetic trees with n tips and $N = 2n-1$ nodes, the range is from $2\log(n)$ to n.

The average shortest path length of a tree may also be informative. Unlike for general graphs, it can be calculated in linear time in a tree with N nodes using dynamic programming [123]. The sum of all shortest path lengths between pairs of nodes in a tree is known as the tree’s Wiener index, and equals $(2n-1)(n-1)$ times the average shortest path length. The Wiener index for general trees with N nodes is always contained between $(N-1)^2$ and $\binom{N+1}{3}$, values which are attained by the star and the path, respectively [51]. For phylogenetic trees, the range of the Wiener index is substantially narrower $(O(n^2\log(n)$ and $O(n^3))$[51]. Note that both the diameter and Wiener index generalize naturally to trees with arbitrary positive branch lengths. The distributions of these indices for small phylogenetic trees, indicates that trees with a diameter or a Wiener index much smaller than the mean are similar to complete binary trees, and those with a diameter or a Wiener index much larger than the mean are similar to caterpillars or double caterpillars, respectively.

Betweenness Centrality

Betweenness centrality associates to each node v in a graph the number of pairs $u, w \in V - \{v\}$ such that the shortest $u-w$ path passes through v; in other words,

$$C_B(v) := \left|\{(u,w) \in V - \{v\} \mid d(u,w) = d(u,v) + d(v,w)\}\right|.$$

Betweenness centrality can be normalized by the number $\binom{N-1}{2}$ of all pairs $u, w \in V - \{v\}$, but we choose not to use this normalization here. In a tree T, there is a unique shortest path between every pair of nodes, and the shortest $u-w$ path passes through v if and only if u and w are in clades subtended by different children of v when T is rerooted at v. Hence, the betweenness centrality of an internal node v is simply $\Pi_{i<j} n_i n_j$, where $n_1, \ldots, n_k$ are the sizes of the clades subtended by the k children of v when T is rerooted at v, and k is the degree of v in T. This implies that the betweenness centrality of all the nodes in a tree can be computed in linear time, a result that, while not surprising, does not seem to have been mentioned in the literature until now.

In a phylogenetic tree, the degree is 1 for a tip, 2 for the root and 3 for internal nodes, so the betweenness centrality is respectively $0, n_1 n_2$ or $n_0 n_1 + n_0 n_2 + n_1 n_2$ in those cases, where n_1 and n_2 are the sizes of the left and right subtrees of the node and n_0 is the number of nodes outside the subtree rooted at this node. This can easily seen to be maximal when $n_0 = n_1 = n_2$, a situation that does not occur in every tree, but is only possible in those with

$N \equiv n \equiv 1 \mod 3$. The betweenness centralities are shown for all the nodes of an example tree in Figure 4.2b. Trees with high maximum betweenness centrality are those which have a node whose left subtree, right subtree, and "outside subtree" (what remains of the tree after removing the subtree rooted at the node), are all close in size - a kind of three-way symmetry, as opposed to the two-way (left-right) symmetry measured by classical statistics

Closeness Centrality

Closeness centrality associates to each node v in a graph the inverse of the sum of its distances to all the other nodes in the graph. In other words,

$$C_C(v) := \frac{1}{\sum_u d(u,v)}.$$

The definition means that closeness centrality is inversely proportional to *farness*, the sum of distances from a node to all the other nodes in the graph; hence, it will be small for centrally located nodes and large for remote ones. While this quantity generally requires at least $O(NM)$ time to be computed for a graph with N nodes and M edges [15], this can be reduced to linear time for a tree, an observation that does not seem to have been published previously, although a distributed algorithm for performing this computation has been proposed [195].

Indeed, if we consider an internal node u with a child v, we have $d(v,x) = d(u,x) - 1$ for every node x in the clade of T that v subtends, and $d(v,y) = d(u,y) + 1$ for every node y outside this clade. Hence, by computing the height (distance to the root) of each node and the sizes of the left and right subtrees of each node in a bottom-up traversal of the tree, we can also find the closeness centrality in linear time for all the nodes in the tree. This is illustrated on our example tree in Figure 4.2c.

Eigenvector Centrality

The Eigenvector centrality $e(v)$ of a node v in a connected weighted graph is defined such that:

$$\sum_{u \sim v} w(uv)e(u) = \lambda e(v)$$

holds for all nodes simultaneously with the largest possible λ. It can be easily seen from this definition that $\vec{e}$ is the eigenvector corresponds to the largest eigenvalue of the graph's adjacency matrix. Using the largest eigenvalue guarantees that the entries of the corresponding eigenvector would be positive. The Eigenvector centrality of all nodes of a phylogenetic tree can be computed in O(n) time since the adjacency matrix of a phylogenetic tree contains exactly $4n - 4$ non-zero entries.

In particular, betweenness centrality, closeness centrality and eigenvector centrality are quantities derived from network science that can be computed in linear time for a tree with

N nodes and can capture aspects of tree shape not captured by mere asymmetry. The tree shape statistics induced by Betweenness centrality, Closeness centrality, and Eigenvector centrality are defined as the maximum values of each centrality over all the nodes of a phylogenetic tree, but using other derived statistics (the minimum, mean, median or variance) could also be an option. Figure 4.2 shows a sample phylogeny with the values of its network statistics.

4.2.3 Spectral Properties Derived from the Distance Laplacian Matrix

In a recent paper, Lewitus and Morlon [105] used the spectra of the distance Laplacian matrix, obtained by subtracting the distance matrix from a diagonal matrix formed by its row sums, to characterize and compare trees. They found promising results and tentatively stated that these spectra were likely to distinguish trees from one another.

We use the four summary statistics proposed by the authors [105] - namely, the maximum eigenvalue, and the asymmetry, kurtosis, and maximum density of the eigenvalue distribution (obtained via smoothing with a Gaussian kernel), as implemented in the *RPANDA* [125] package in R [154]. We note that, unlike the network statistics, these ones require a number of operations proportional to n^3, where n is the number of tips, and thus take substantially longer to compute.

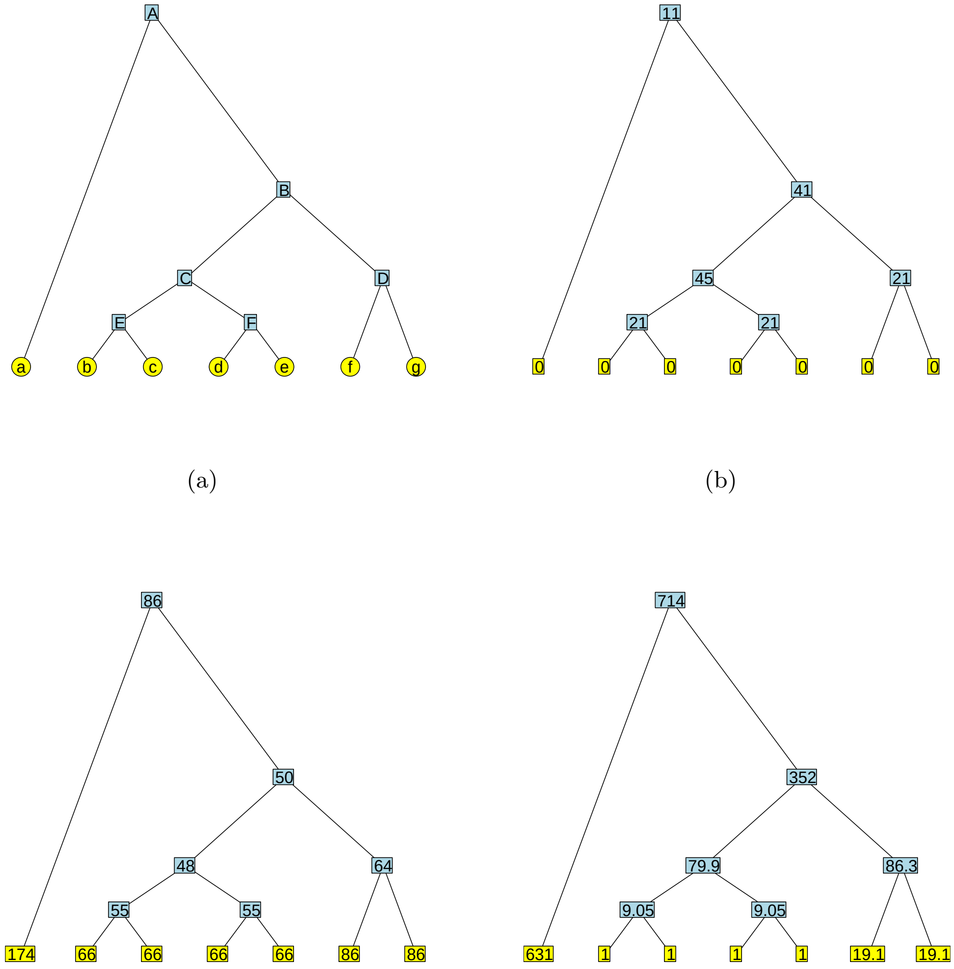
4.2.4 Data and Simulations

In order to evaluate the power of network statistics in distinguishing between trees, we applied them along with conventional tree statistics to phylogenetic trees simulated using different parameters and reconstructed from different viruses.

HIV/Dengue/Measles We obtained Newick tree strings corresponding to phylogenies inferred from human and zoonotic RNA viruses from a previous study. Specifically, we retrieved tree strings reconstructed from genetic sequences of HIV-1 subtype B, Measles virus, and Dengue virus serotype 4. The HIV sequence data (corresponding to the gene encoding the Nef protein) were obtained from the LANL HIV Sequence database [59], through the web site at *http://www.hiv.lanl.gov*, and screened for recombinants using the SCUEAL algorithm [96]. The remaining virus sequences were obtained from GenBank [12].

Phylogenies were reconstructed from random samples of 100 sequences by maximum likelihood using RAxML [181] under a general time-reversible model of nucleotide substitution and rate variation among sites approximated by the GTRCAT model. HIV-1 subtype B phylogenies were rooted using a subtype D sequence as an outgroup. Dengue virus serotype 4 phylogenies were rooted on an outgroup sequence isolated in the Philippines in 1956. Finally, measles virus phylogenies were rooted using a genotype D6 sequence as the outgroup. The GenBank [12] accession numbers for all outgroups can be found in Table A.1.

HIV in three settings We obtained HIV-1 sequence data from three published studies. The Wolf et al. [200] data set corresponds to samples from a concentrated epidemic of HIV-1



(a)

(b)

(c)

(d)

Figure 4.2: A sample phylogenetic tree with the associated centrality values at each node. (a) An example tree. In the epidemiological context, tips $(a - g)$ would correspond to pathogen sequences and internal nodes $(A - F)$ to their inferred common ancestors. Node D subtends a "cherry" configuration, and node C subtends two cherries (a "double cherry"). The heights of the internal nodes are 1 (E, F), 2 (C, D), 4 (B) and 8 (A), so the diameter is 16 and the Wiener index is 484, for a mean path length of 6.21. (b) Same tree with betweenness centrality values at each node (note that branch lengths do not change them). The tree has betwenness centrality 45. (c) Same tree with farness (reciprocal of closeness centrality) values at each node. The tree has closeness centrality 1/48. (d) Same tree with eigenvector centrality values (scaled to have a minimum of 1) at each node, rounded to 3 significant figures. Here, the leading eigenvalue is $\lambda = 9.05$. By our definition, the tree has eigenvector centrality $714/1023 = 0.698$ (here, 1023^2 is the sum of the squared values).

subtype B in populations of predominantly men who have sex with men in Seattle, USA. The Novitsky et al. [137] data set corresponds to samples from a generalized epidemic of HIV-1 subtype C infections in Mochudi, Botswana, a village with an estimated HIV-1 prevalence of about 20% in the adult population. Similarly, the Hunt et al. [81] data set represents samples from a national survey of the generalized epidemic of HIV-1 subtype C in South Africa. Thus, these studies represent a range of geographic scales and epidemiological contexts.

Influenza in three settings We compared the topologies of a single virus (human influenza A) sampled to reflect different epidemiology: (1) samples over a five-year period; (2) global samples over a 12-year period and (3) samples from 2012-2013 from the USA only. We downloaded full-length hemagglutinin (HA) sequences of human H3N2 flu from NCBI and aligned sequences from each group with MAFFT [88]. For each sample, we chose 120 sequences uniformly at random from the alignment, and inferred a tree with these sequences as tips using IQtree [191] with the pll (phylogenetic likelihood library) option [58] and the GTR+G model. Using the date information from NCBI, we rooted the trees using the root to tip (*rtt*) function in the *ape* package [140] in R [154].

Simulated tree models We simulated trees from four random processes: a Yule process (pure birth trees), a "biased" model of Kirkpatric and Slatkin [93] in which speciation rates are unevenly assigned to a node's descendants with a bias (here 0.3), and a two constant rate birth-death processes, with the basic reproduction number (mean of the offspring distribution) equal to 1.5 and 3. We created sets of 100 trees with 100 tips and separately with 300 tips. The *apTreeshape* package was used to simulate the Yule and biased models; tree shapes were converted to phylogenetic trees using the *as.phylo* function. We used the *TreeSim* package for the birth-death models. Because in *sim.bd.taxa* (in *TreeSim*) the simulations are conditioned on having a fixed number of extant tips, we created trees with 300 or 600 extant tips and randomly pruned taxa to leave a tree of 100 or 300 tips, modelling partial sampling over time.

As some scenarios can be distinguished simply by comparing the branch lengths of the corresponding trees, we normalized the time scales so that each of our trees has a mean branch length of 1. This ensures that any differences we observe between the summary statistics in different classes are not simply due to scaling. We did not, however, modify the variances of the branch length distribution, as those may contain some of the signal picked up by summary statistics.

In total, there are 5 scenarios in which we compare trees: HIV/Dengue/Measles (HDM), influenza (2-year USA, 5-year global, 12-year global), HIV contexts (labeled WNH after the first author names of the corresponding publications), simulated trees with 100 tips ("Simulated") and simulated trees with 300 tips ("Simulated300"). Within each set, we performed classification with generalized linear models and random forests, using the tree shape statistics as features. We computed a measure of each feature's importance for each classification.

4.3 Results

In this section, we elaborate the experiments we perform to evaluate the power of our suggested statistics.

4.3.1 Comparing the Power of Saless with the Classical Statistics

In this part, we experiment with some of the classical statistics introduced in chapter 2 and our suggested statistic *Saless*. Table 4.2 represents the scaled resolution scores based on the distance resolution function for comparison of the various statistics. Each row in the table contains the resolution for trees with a given number of leaves, while each column contains the resolution for each statistic.

n	I	$\bar{S}$	σ_n^2	I_2	B_1	B_2	Saless
7	0.925	0.93	0.902	0.884	0.865	0.925	0.931
8	0.926	0.912	0.875	0.861	0.833	0.911	0.926
9	0.918	0.921	0.883	0.854	0.832	0.907	0.923
10	0.941	0.938	0.898	0.855	0.833	0.908	0.943
11	0.953	0.951	0.91	0.855	0.837	0.913	0.955
12	0.953	0.952	0.909	0.85	0.831	0.904	0.956
13	0.954	0.954	0.908	0.842	0.825	0.899	0.957
14	0.955	0.955	0.907	0.837	0.82	0.89	0.957
15	0.955	0.954	0.905	0.83	0.813	0.883	0.956
16	0.954	0.954	0.903	0.827	0.809	0.874	0.956
17	0.953	0.953	0.901	0.82	0.802	0.868	0.956

Table 4.2: Scaled resolution scores for tree statistics based on the distance resolution function on the *NNI* distance matrix. The resolution is between 0 and 1. n is the number of leaves. The tree shape statistics are described in chapter 2. The highlighted values correspond to the best statistics in each row.

As this table shows, our proposed statistic has higher resolution than the previously defined ones.

4.3.2 Linear Combination of the Classical Tree Shape Statistics

We investigate pairwise linear combinations of statistics based on our new resolution function (Laplacian resolution function). The linear combination of the Colless index and B_2 performs better than all other statistics. Similarly, the linear combination of the Colless and Sackin indices results in a high resolution. The results of this experiment are summarized in Table 4.3.

n	$I-B_2$	λ	B_2-B_1	λ	Saless	λ
7	0.0922	-0.08	0.0855	2.89	0.0932	0.08
8	0.0799	-0.28	0.0884	3.43	0.076	-1.2
9	0.0505	-0.53	0.0576	3.2	0.0502	-2.36
10	0.0324	-0.3	0.0405	4.14	0.0323	-2.49
11	0.0221	-0.5	0.0306	4.22	0.0221	-4.3
12	0.0182	-0.49	0.0273	4.91	0.0181	-2.87
13	0.0160	-1.4	0.0256	5.14	0.0159	-3.56
14	0.0147	-3.58	0.0244	5.69	0.0146	-3.19
15	0.0137	1.31	0.0237	6.03	0.0136	-3.17
16	0.0129	0.67	0.0230	6.5	0.0128	-2.8
17	0.0123	0.41	0.0224	6.86	0.0122	-2.69
18	0.0116	0.3	0.0217	7.28	0.0115	-2.52
19	0.0111	0.24	0.0212	7.65	0.0110	-2.4
20	0.0105	0.2	0.0206	8.04	0.0105	-2.27
21	0.0100	0.17	0.0200	8.4	0.0100	-2.17
22	0.0096	0.14	0.0195	8.77	0.0096	-2.08
23	0.0092	0.13	0.0190	9.12	0.0092	-2.00
24	0.0088	0.11	0.0185	9.47	0.0088	-1.94
25	0.0084	0.10	0.0180	9.82	0.0084	-1.88

Table 4.3: Scaled resolution scores for the optimal linear combinations of $I-B_2$, B_2-B_1, and *Saless* based on our new proposed resolution function (Laplacian resolution function). The corresponding optimal values of λ are shown next to each combination.

4.3.3 Network Statistics Improve Classification of the Trees Reconstructed From Different Viruses and Epidemiological Scenarios

We computed a range of topological and spectral summary features of the viral phylogenies (see Table 4.4 for the definitions and references for each one). Our focus here is on some of the novel tree shape summary statistics, but we also include a number of standard statistics for comparison. All input trees were binary and rooted, and all branch lengths were non-negative, although many of the trees had zero-length branches. All comparisons involved trees on the same number of tips.

For the node properties derived from network science, we focus our discussion on the maximum value of each type of centrality a node can have within a tree, but using other derived statistics (the minimum, mean, median or variance) could also have been an option. For the spectral properties, which also produce a value for each node, we focus on the maximum value as well as the minimum strictly positive value. Lastly, for the distance Laplacian spectral properties we use the four derived statistics proposed by Lewitus and Morlon [105].

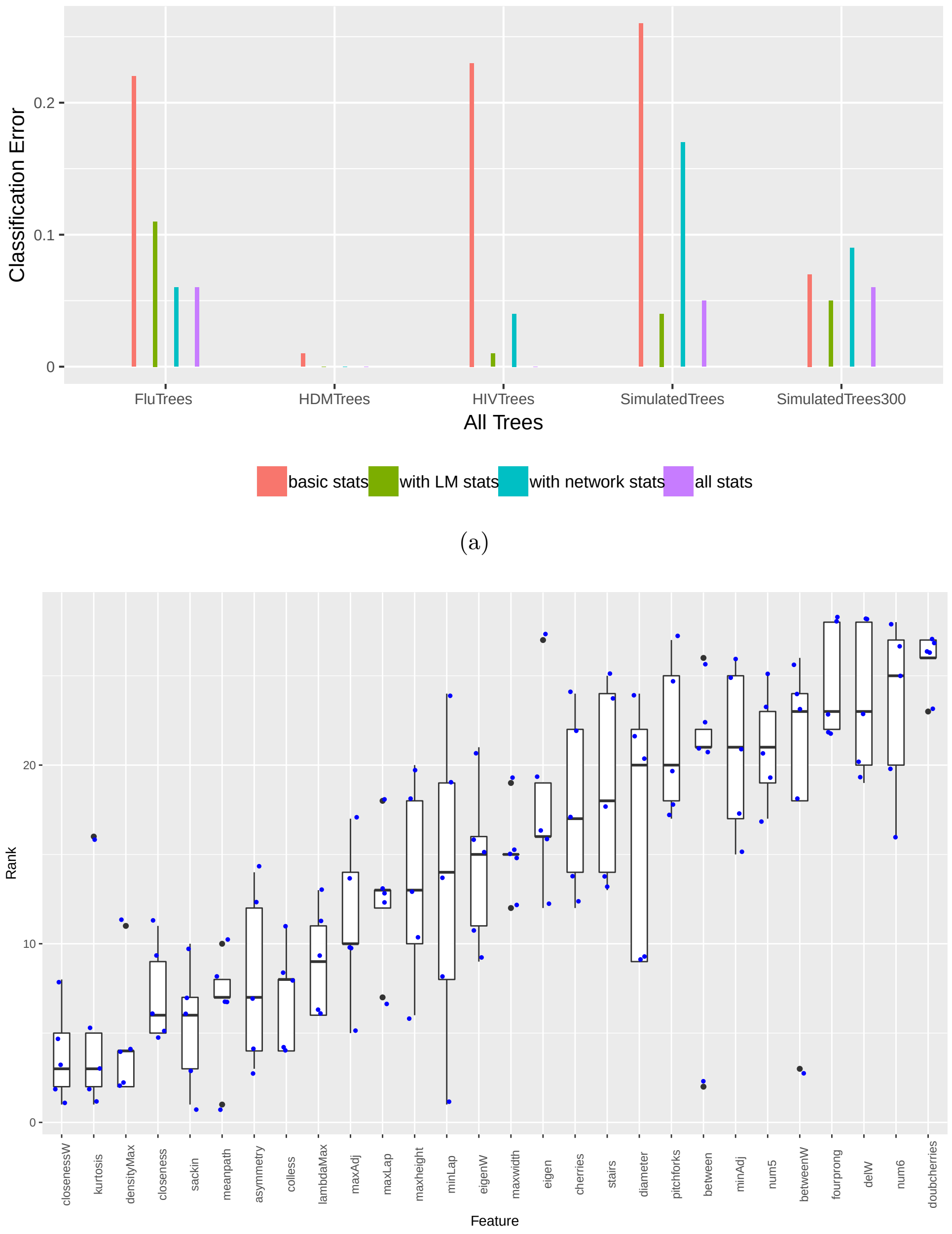
We use a generalised linear model and two flavours of random forests to classify trees within each scenario. For example, we classified trees as HIV, Dengue or Measles; we

Name	Description	Short form	Ref.
Numbers of small configurations			
Cherry number	# of nodes with 2 tip children	cherries	[118]
Pitchforks	# of nodes with 3 tip descendants	pitchforks	[162]
Double cherries	# of nodes with 2 cherry children	doubcherries	new
4-caterpillar	# of caterpillars with 4 tips	fourprong	[162]
Clades of size x	# of nodes with x tip descendants	num$x$	[162]
Tree-wide summaries			
Colless imbalance	Colless imbalance	colless	[34]
Sackin imbalance	Mean path length from tip to root	sackin	[163]
Maximum height	Max # of steps from the root	maxheight	[32]
Maximum width	Max # of nodes at the same depth	maxwidth	[32]
Stairs	Proportion of imbalanced subtrees	stairs	[136]
Max difference in widths	$\max_i(n_{i+1} - n_i)$	delW	[32]
Node properties from network science			
Betweenness centrality	# of shortest paths through node	between	[133]
Weighted betweenness	as above, but with weighted edges	betweenW	[133]
Closeness centrality	total distance to all other nodes	closeness	[133]
Weighted closeness	as above, but with weighted edges	closenessW	[132]
Eigenvector centrality	value in Perron-Frobenius vector	eigen	[133]
Weighted eigenvector	as above, but with weighted edges	eigenW	[132]
Summaries from network science			
Diameter	largest distance between 2 nodes	diameter	[19]
Mean pairwise distance	average distance between 2 nodes	meanpath	[123]
Spectral properties			
Min adjacency	min adjacency matrix eigenvalue > 0	minAdj	[65]
Max adjacency	max adjacency matrix eigenvalue	maxAdj	[65]
Min Laplacian	min Laplacian matrix eigenvalue > 0	minLap	[65]
Max Laplacian	max Laplacian matrix eigenvalue	maxLap	[65]
Distance Laplacian spectral properties			
Max eigenvalue	largest eigenvalue in the spectrum	dLapLambdaMax	[105]
Max density	location of largest spectral density	dLapDensityMax	[105]
Asymmetry	skewness of the spectral density	dLapAsymmetry	[105]
Kurtosis	peakedness of the spectral density	dLapKurtosis	[105]

Table 4.4: Summary measures for phylogenetic trees. Here, n_i is the number of nodes at depth i.

classified influenza trees as five-year, global, and USA; we classified simulated trees as biased, Yule, $R_0 = 1.5$ or 3; and we classified the HIV trees by epidemiological scenario. In each classification task, we randomly selected 75% of the trees (75 trees from each group) for training and used the remaining 25 trees from each group for testing, and computed the classification error of the predictor on the testing set. We report the overall classification error with and without the features based on network science, as well as with and without the features based on the distance Laplacian spectrum (abbreviated as "LM" statistics after the authors Lewitus and Morlon). The results are shown in Figure 4.3(a). It is clear that the standard ("basic") tree shape statistics are not able to get close to perfect classification on any of the datasets except for the (relatively simple) task of distinguishing three different viruses. The addition of the costly LM statistics improves the performance, but so does the addition of the easily computable new network science-based statistics we introduced in this chapter, with comparable gains in performance relative to the baseline (a larger improvement on the flu scenarios, and a slightly smaller one on the HIV scenarios and on simulated data). Interestingly, the addition of network statistics actually increases the error relative to the baseline on the large simulated trees, an effect that is likely due to the random nature of the classifier.

Both random forests and the GLM regression additionally provide estimates of the importance of each feature. For the random forest classifier, the prediction accuracy on the out-of-bag portion of the data is recorded for each tree, and then the same is done after randomly permuting each predictor variable. The differences between the two accuracies are then averaged over all trees, and normalized by the standard error. For the GLM regression, the absolute value of the t-statistic for each model parameter is used as the importance measure. We used the *caret* package [99] in R [154] to compute these for all the classifiers. For each classification task in each scenario, we then ranked the features by importance. We show the rank data in Figure 4.3(b). It is apparent that weighted closeness centrality is the feature with the highest importance (lowest rank) across all scenarios, on average, with the much costlier to compute distance Laplacian-based features coming close behind.



(a)

(b)

Figure 4.3: (a) Accuracy of classification with and without network science features, as well as with and without the distance Laplacian spectral features. (b) Feature importance in multi-class classification across all scenarios, ordered by the median. Each point is the rank of the corresponding feature in one of the classification tasks. Low ranks correspond to the most important features (i.e. the top-ranked feature has rank 1, and so on).

Chapter 5

An Application of Tree Shape Statistics

Human influenza A virus remains a substantial global public health challenge, causing considerable illness and mortality despite the availability of effective vaccines. Influenza viruses are categorized according to features of two surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA), with types such as H3N2 and H1N1 indicating the variant of HA and NA characterizing the strain. Influenza viruses are prone to variability, both in the form of so-called antigenic drift (small mutations in the genes of influenza viruses that can lead to changes in the surface proteins of the virus: HA and NA), and in the form of reassortment (the process by which influenza viruses swap gene segments) [174]. Reassortments can give rise to new variants with distinct antigenic properties compared to previous strains; resulting pandemic influenza virus strains may be highly pathogenic. In contrast to pandemic strains arising from reassortment, seasonal influenza virus primarily arises through antigenic drift, as influenza virus has a high propensity for generating antigenic variation [174]. This allows influenza viruses to evade host population immunity built up through previous exposure. As a consequence, seasonal influenza virus vaccines need to be regularly updated.

Influenza virus vaccines typically focus on preventing infection by raising antibodies specific to the hemagglutinin (HA) protein. In order to update a seasonal influenza virus vaccine, currently-circulating strains must be selected for inclusion. This relies on surveillance and sequencing of circulating influenza virus genotypes, and on measured antigenic properties of circulating strains. These data do not, in and of themselves, describe future circulating strains, and sometimes the strain selection process does not reflect the near-future composition of the influenza viruses well enough to achieve the desired reductions in illness and mortality. Predictive models are now being used in conjunction with sequencing and immunological surveillance in order to improve the strain selection process.

Phylogenetic trees have been used in infectious disease to estimate the basic reproduction number [179], parameters of transmission models [194], aspects of underlying contact networks [146, 102, 158, 119] and in densely sampled datasets even person-to-person transmission events and timing [46, 94, 9, 202]. It is therefore natural to hypothesize that phylogenetic tree structures and branching patterns contain information about short-term growth and fitness. Tree information is central in some predictive models for short-term influenza virus evolution and models of fitness [131, 43]. However, the mapping between the phylogenetic tree structure and interpretable biological information can be subtle, [158, 119, 31, 117] and trees do not directly reveal the short-term evolutionary trajectories of groups of taxa.

Improvements in influenza virus surveillance, sequencing, data sharing and visualization [129, 72] mean that sequence data over considerable time frames is now available to the community alongside intuitive and interactive displays showing how the population of influenza viruses has changed over time. Computational systems to reconstruct large-scale phylogenetic trees from sequence data have also been developed [181, 21, 25, 148]. Machine learning models are well-suited to systematically explore subtle relationships between a suite of features and an outcome. These together present the opportunity to integrate information from different sources to improve short-term influenza virus predictions, using phylogenetic trees as a framework. Here, we use a convolution-like approach to identify small subtrees within a large global H3N2 phylogeny derived from HA sequences sampled between 1980 and May 2018. We fit classification models to detect early signs of growth and hence to predict the short-term success or failure of these subtrees. We validate the predictions on a portion of the data not included in the fitting procedure. We relate our predictions to the WHO defined clades [10, 11] for sequences sampled from 2015-2018. Our approach could be performed in real time, is computationally efficient and can be continually refined to improve the quality of predictions as more data are gathered. We suggest that small groups of closely-related influenza virus sequences and the phylogenetic trees that capture their recent shared ancestry patterns can complement other approaches to better predict short-term seasonal influenza virus evolution.

5.1 Problem Definition

Seasonal influenza viruses are constantly changing, and produce a different set of circulating strains each season. Small genetic changes can accumulate over time and result in antigenically different viruses; this may prevent the body's immune system from recognizing those viruses. Due to rapid mutations, in particular in the hemagglutinin gene, seasonal influenza vaccines must be updated frequently. This requires choosing strains to include in the updates to maximize the vaccines' benefits, according to estimates of which strains will be circulating in upcoming seasons. This is a challenging prediction task. Here, we use longitudinally sampled phylogenetic trees based on hemagglutinin sequences from human influenza viruses, together with counts of epitope site polymorphisms in hemagglutinin, to predict which influenza virus strains are likely to be successful. In this chapter, we consider labeled rooted phylogenetic trees.

5.2 Methods

Our approach is rooted in the hypothesis that fitness – the reproductive rate and capacity of a group of organisms – affects tree structure and branching patterns (including timing) and that this information can be extracted using machine learning tools.

5.2.1 Reconstructing Influenza Virus Trees

We collected full-length HA sequences from human H3N2, H1N1 and influenza virus B from GenBank [13, 168]. We used unique sequences of H3N2-HA from human cases for years between 1980 and 2018-5 (May 2018), excluding laboratory strains. This results in approximately 12919 sequences. For influenza virus H1N1 we collected pandemic sequences from 2009 until 2018-5, giving 10652 unique sequences. For influenza virus B, we used sequences from 1980 to 2018-5, resulting in 7257 unique sequences. We aligned each set of sequences using MAFFT [87, 88] and then we used RAxML [181] to reconstruct maximum-likelihood phylogenetic trees. The reconstructed trees using RAxML are neither rooted nor dated; in some cases they include very long branches (edges) in comparison with the mean branch length, which we removed. We rooted trees with the *rtt* function in the *ape* package [140] in *R* and then converted them to timed trees using the Least Squares Dating (*LSD*) software [189].

5.2.2 Subtree Extraction

We use a convolution-style approach to identify subtrees of the global timed phylogeny that serve as units of analysis. For each internal node i in the tree, we find the tips that occur within a fixed time window (1.4 years by default) chronologically following i; this is node i's "trimmed subtree". We cannot train machine learning models on the subtree descending from every internal node in the tree, because these subtrees will overlap substantially. We use the notion of a node's "relevant ancestor" (described below) to control the overlap, and select subtrees in a convolution-like way.

We first initialize each node's relevant ancestor to be its parent. We traverse the nodes of the tree in a depth-first search order. If a node's complete subtree is too small (fewer than 8 nodes by default), we reject the node and all its descendants, as none of the descending nodes can have a larger subtree than the node itself. If the node's trimmed subtree is too small but its complete subtree is large enough, we reject the node but not its descendants, since they may have subtrees that are large enough. If the node's trimmed subtree is large enough, we check the overlap between the node's subtree and its relevant ancestor's subtree. The overlap is the portion of node i's trimmed subtree that is contained in the relevant ancestor's trimmed subtree. If this overlap is not too large (under 80% of the subtree size by default), the subtree is included in our analysis. If the overlap is too large, we reject the node, and we set the relevant ancestor of the node's children to be the node's relevant ancestor. In that way, when we decide whether to accept the subtree of the node's child, we will control the correct overlap (see Algorithm 1).

Algorithm 1 takes an influenza tree, minimum accepted size for the subtrees (*min_total_size*), minimum accepted size for the trimmed subtrees (*min_trim_size*), minimum allowed overlap between the trimmed subtrees ($overlap\_cutoff$), and the time that we trim the subtrees after that time from the root of the subtrees (*time_frame*) as inputs. The algorithm outputs the set of accepted nodes with the tips in the trimmed subtrees. Line 4 computes the number of tips in the input tree and store it in *num_tips* Line 6 stores the Depth First Search traversal order of the nodes of the input tree in *internal_nodes* array. Line 7 creates an array (*relevant_parent*) of size equal to the number of internal nodes. Lines 8 and 9 initialize the relevent parent of each node. Line 10 creates a Boolean array of size equal to the number of internal nodes in the tree and initiates each element to False. Line 11 stores the first element of the *internal_node* as root. Line 12 changes the corresponding entry of the root in the $reject\_flag$ array to True. The **for** loop of lines $13 - 33$ is the main part of the algorithm with the following loop invariant:

At the start of each iteration of the **for** loop *current_subtree_root* contains the root of the current subtree under consideration. The variables $left\_child$ and *right_child* contain the left and right child of *current_subtree_root*. The variables $index\_left\_child$ and *index_right_child* contain the indices of $left\_child$ and *right_child* in the *internal_node* array.

Line 17 checks if the $reject\_flag$ of the root of the current subtree is not True. If the condition satisfies then Line 18 retrieves the relevant parent of the current subtree root from the *releveant_parent* array. Line 19 computes the number of tips in the current subtree under consideration. Line 20 first trims the nodes in the current subtree with depth greater than $time\_frame$ and then computes the number of tips in the trimmed subtree. Line 21 checks if the size of the current subtree is less than minimum allowed size. If this condition is true, then we ignore this subtree and all of its descendent from future consideration by making their $reject\_flag$ True in Lines $22 - 24$. Line 25 checks if the size of the current subtree is greater than the minimum total allowed size and the number of tips in the trimmed subtree is less than the minimum allowed size of the trimmed subtree. If this condition is true, then we ignore only the root of the current subtree by making its $reject\_flag$ True in Line 26. Line 27 checks if the size of the current subtree is greater than the minimum total allowed size and the number of tips in the trimmed subtree is greater than the minimum allowed size of the trimmed subtree. If this condition is true, then Line 28 computes the number of tips that share between the the trimmed subtree rooted at current node and the trimmed subtree rooted at the relevent parent of the node. Line 29 checks if the number of shared nodes between these two subtrees is greater than the minimum overlap threshold (defined by $overlap\_cutoff \times num\_tips\_trimmed$) then we ignore the current root by making its $reject\_flag$ True in line 30. It also updates the relevant parent of the children of the current root to be the relevant parent of the current root in lines $31-32$. After the loop of line $13-32$, the algorithm retrieves those nodes with $reject\_flag$ equal to False as the accepted nodes in line 33. Then it outputs these accepted nodes together with the tips in the trimmed subtree rooted at these accepted nodes in line 34.

In this way, we obtain subtrees containing tips that are within a specified time window after their originating node, have at least a minimum number of tips, and have a limited overlap with other subtrees. We varied the minimum size, time interval and permitted overlap (See Table 5.2). We obtain a total of 396 subtrees in H3N2, and 198 subtrees in H1N1. After removing subtrees with large size in relation to the average size (approximately greater than 10 times the average size), and recent subtrees with insufficient growth to determine their outcome, we obtained 329 subtrees for H3N2 and 160 subtrees for H1N1.

5.2.3 Features

We use a set of features defined on subtrees, including both tree shape and patterns in the branch lengths. The topological features are summarized in Table 5.1. For the H3N2-HA dataset, we also consider some features derived from the epitope sites of the tips of the subtree. For each subtree, we consider the mean, median and maximum genetic distances between the epitope sites of the tips of a subtree and the epitope sites of the sequences with dates prior to the subtree. We used the locations of known antigenic epitopes as mentioned in [175], namely 72 sites in the *HA1* subunit of HA.

Our features cover a wide range of global and local structures in trees, expanding considerably over previous approaches which largely focus on tree asymmetry and a few properties of branch lengths [131, 43]. Previous authors have noted that fitness leaves traces in genealogical trees [43] by observing in fixed-size populations that increased fitness resulted in increased asymmetric branching and long terminal branch lengths; Neher and colleagues used the local branching index (LBI), a measure of the total branch length surrounding a node, in their predictive model [131]. We significantly expand on the repertoire of tree features, including asymmetry and measures of local branching but also including features derived from network science that capture global structure of the subtrees, small shape frequencies and others – see Table 5.1.

For comparison purposes, we implemented Neher et al.'s local branching index (LBI) [131], which is a measure of rapid branching near a node in the tree. In doing this, we noted that there are strong parallels between the LBI and the weighted version of the Katz centrality, a classic measure from network science [89]. Figure B.2 shows the correspondence. We performed the main classification task (H3N2) using all our features, only the topological tree features, only the epitope and LBI, only the epitopes and only the LBI (Figure 5.5(b)). We found that the combined features gave the best performance, followed by the tree features.

Algorithm 1 Subtree extraction algorithm

```
Input: tree, $min\_total\_size$, $min\_trim\_size$ , $overlap\_cutoff$, $time\_frame$
Output: A list of nodes and the tips in the nodes trimmed subtrees
function GETSUBTREES(tree, $min\_total\_size$, $min\_trim\_size$, $overlap\_cutoff$, $time\_frame$)    ▷ the tree should be rooted and timed
    $num\_tips \leftarrow$ number of tips in the tree
    $num\_internal\_nodes \leftarrow num\_tips - 1$
    Let $internal\_nodes$ be an array storing the internal nodes of the tree in a DFS traversal (pre-order traversal) of the tree
    Let $relevant\_parent$ be an array of size equal to $num\_internal\_nodes$
    for i in 1: $num\_internal\_nodes$ do
        $relevant\_parent[i] = parent[internal\_nodes[i]]$    ▷ parent of the root node is itself
    Let $reject\_flag$ be an array of size $num\_internal\_nodes$ initiated to $False$
    root $\leftarrow internal\_nodes[1]$
    $reject\_flag[root] \leftarrow True$
    for k in 2: $num\_internal\_nodes$ do
        $current\_subtree\_root \leftarrow internal\_nodes[k]$
        $left\_child, right\_child \leftarrow children(current\_subtree\_root)$
        Let $index\_left\_child, index\_right\_child$ be the indices of the $left\_child$ and $right\_child$ in $internal\_nodes$ respectively.
        if $reject\_flag[k] \neq True$ then
            $relevant\_parent\_current\_root \leftarrow relevant\_parent[k]$
            $num\_tips\_current\_subtree \leftarrow$ the number of tips in the subtree rooted at $current\_subtree\_root$
            $num\_tips\_trimmed \leftarrow$ the number of tips in the trimmed subtree rooted at $current\_subtree\_root$ (trim the nodes with depth greater than the $time\_frame$)
            if $num\_tips\_current\_subtree < min\_total\_size$ then
                $reject\_flag[k] \leftarrow True$
                $reject\_flag[index\_right\_child] \leftarrow True$
                $reject\_flag[index\_left\_child] \leftarrow True$
            else if $num\_tips\_current\_subtree \geq min\_total\_size$ and $num\_tips\_trimmed < min\_trim\_size$ then
                $reject\_flag[k] \leftarrow True$
            else if $num\_tips\_current\_subtree \geq min\_total\_size$ and $num\_tips\_trimmed \geq min\_trim\_size$ then
                intersect $\leftarrow$ the number of tips that are shared between the trimmed subtree rooted at current node and the trimmed subtree rooted at the relevent parent of the node
                if intersect $> overlap\_cutoff \times num\_tips\_trimmed$ then
                    $reject\_flag[k] \leftarrow True$
                    $relevant\_parent[index\_left\_child] \leftarrow relevant\_parent[k]$
                    $relevant\_parent[index\_right\_child] \leftarrow relevant\_parent[k]$
    $accepted\_subtrees \leftarrow$ the $internal\_nodes$ with $reject\_flag$ equal to $False$
    return $accepted\_subtrees$ and the tips in the trimmed subtrees rooted at $accepted\_subtrees$
```

Name	Description	Reference
Properties from network science		
Betweenness centrality	Max number of shortest paths through nodes	[133]
Closeness centrality	Max total distance to all other nodes	[133]
Eigenvector centrality	Max value in Perron-Frobenius vector	[133]
Diameter	Largest distance between 2 nodes	[19]
WienerIndex	Sum of all distances between 2 nodes	[123]
Mean tips pairwise distance	Average distance between 2 tips	natural
Max tips pairwise distance	Max distance between 2 tips (with branch lengths)	weighted graph diameter
Numbers of small configurations		
Cherry number	Number of nodes with 2 tip children	[118]
Normalized Pitchforks	3*(Number of nodes with 3 tip descendants) / n	[162]
Tree-wide summaries		
Normalized Colless imbalance	$\frac{1}{n^{3/2}} \sum_{i \in \mathcal{I}} \lvert r_i - s_i \rvert$	[35]
Normalized Sackin imbalance	$\frac{1}{n^{3/2}} \sum_{i \in \mathcal{L}} N_i$	[163]
Normalized Maximum height	The maximum height of tips in the tree. / $(n-1)$	[91]
Maximum width	Max number of nodes at the same depth	[31]
Stairs1	The portion of imbalanced subtrees	[136]
Stairs2	The average of $\frac{min(s_i,r_i)}{max(s_i,r_i)}$ over all internal nodes	[136]
Max difference in widths	$\max_i(n_{i+1} - n_i)$	[31]
Variance	The variance of internal node depth	[115]
I2	$\sum_{\substack{i \in \mathcal{I} \cup \{r\} \\ r_i + s_i > 2}} \frac{\lvert r_i - s_i \rvert}{\lvert r_i + s_i - 2 \rvert}$	[115]
B1	$\sum_{i \in \mathcal{I}} M_i^{-1}$	[115]
B2	$\sum_{i \in \mathcal{L}} \frac{N_i}{2^{N_i}}$	[115]
Normalized Average ladder	The mean size of ladders in the tree / $(n-2)$	[91]
Normalized ILnumber	Number of internal nodes with a single tip child / $(n-2)$	[91]
Branching speed	The ratio of the number of tips to the height of the tree	new
Measures from edge length		
Branching next index	Mean of indicator: does the next branching event descend from this node	new
Generalized branching next	Number of next two branching events descending from this node	new
Skewness	The skewness of the internal branch lengths	natural
Kurtosis	The kurtosis of the internal branch lengths	natural

Table 5.1: Brief definition for tree shape statistics. Here r_i and s_i are the number of tips of the left and right subtrees of an internal node respectively. n is the number of tips of a subtree. n_i is the number of nodes at depth i, M_i represents the height of the subtree rooted at an internal node i, and N_i is equal to the depth of node i. A ladder in a tree is a set of consecutive nodes with one tip child. We represent the set of all internal nodes of a tree by $\mathcal{I}$, the set of all tips (or external nodes) by $\mathcal{L}$. In "generalized branching next" we chose $m = 2$. Skewness and Kurtosis are two measures to describe the degree of asymmetry of a distribution [114]. The tree shape statistics induced by betweenness centrality, closeness centrality and eigenvector centrality are defined as the maximum values of each centrality over all the nodes of a tree, and distances are in units of number of edges (without branch lengths). Features called "natural" may not have been used as tree features previously but are natural extensions of simple features (eg skewness is a natural quantity to compute). The network science properties were computed in R using the treeCentrality package [101] and the tree-wide summaries were primarily obtained using the phyloTop package [91]

5.2.4 Success and Training Approach

We call a subtree of size n "successful" if its root has a total of more than αn tip descendants in the time frame of 3.4 years from the root of the subtree. The threshold of $\alpha = 1.1$ results in a good balance of successful and unsuccessful subtrees, which facilitates training the machine learning models (see Figure 5.1). We chose to use fractional growth as our outcome rather than proximity to tips of the following season, because proximity to the following season fluctuates depending on when in the season the subtree originates, the definition of the season (i.e. the cutoff dates) and the subtree's location (tropical vs temperate).

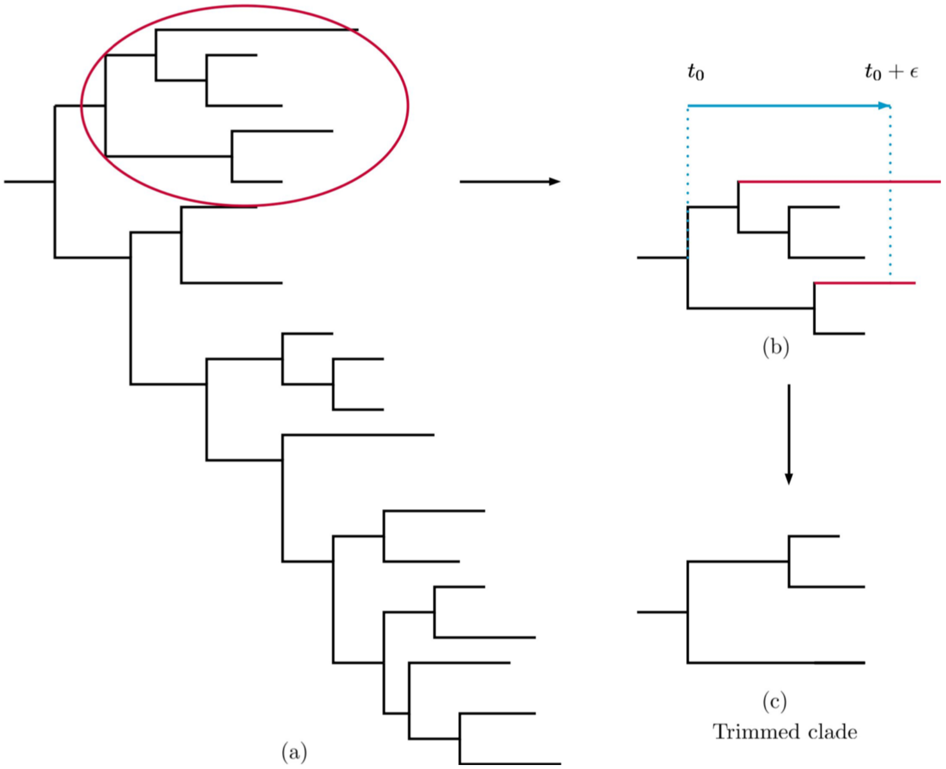


Figure 5.1: Illustration of the formation of trimmed subtrees: (a) the circled clade contains a subtree; (b) red branches reach tips that occur after the trimming time period and so are pruned out; (c) the resulting trimmed subtree.

5.2.5 Classification

We use several different binary classification tools, including support vector machines (SVM) with a range of kernel choices [37]. We use R implementations in the package *e1071* [47]. For all the experiments, we randomly chose 75% of our subtrees for training the model and the rest for testing. We perform 10-fold cross validation on the training dataset alone; this is to find the best gamma and cost parameters (the parameters that are utilized in constructing the SVM) without using all the data to do so (see Figure 5.2). Among different binary classification tools that we used, SVM with a linear kernel had the best performance on this 75% training set, so we proceeded with this option for the remaining results. Datasets can have outliers that affect the training process. In order to remove the outliers, we use the local outlier factor (LOF) algorithm [22] implemented in the *DMwR* package [190] in *R*. Local Outlier Factor is the anomaly score of each data point. It measures the local deviation of density of a given sample with respect to its neighbors. The reason that it is called local is that the anomaly score depends on how isolated the object is with respect to the surrounding neighborhood. More precisely, locality is given by k-nearest neighbors, whose distance is used to estimate the local density. By comparing the local density of a sample to the local densities of its neighbors, one can identify samples that have a substantially lower density than their neighbors. These are considered outliers. For classification on the H3N2 tree, we removed 5 outliers which were mostly large subtrees, most of whose descending tips were contained in other subtrees. In the H1N1 tree, we had 198 subtrees; we removed the largest 11 as outliers, for the classification. For the experiment on the merged H1N1 and H3N2 subtrees we removed large subtrees of both H3N2 and H1N1 to have a set of subtrees of approximately the same size. In the experiments on the influenza virus B tree and the combinations of influenza virus B and other types, we remove large subtrees to obtain a set of trees of approximately the same size in the training and testing groups. Our training and testing scheme is illustrated in Figure 5.2.

5.2.6 Time Slices

To ensure that the method does not rely on internal nodes whose existence depends on the full dataset, we reconstructed the influenza virus H3N2 tree using only the sequences observed prior to time i ($i \in \{2012-5, 2013-5, 2014-5, 2015-5, 2016-5, 2017-5\}$); call this tree T_i. We extracted the subtrees of this tree using the ALT0 parameters (see Table 5.2 for the details of ALT0). Naturally, the T_i tree does not contain the information as to whether its later-occuring subtrees grow into the following season. To find the remaining success information, we used the tree reconstructed from the sequences any time up to $i+1$ (T_{i+1}). Consider a subtree c in T_i, with tip set S_c and size n. First we find the most recent common ancestor (*MRCA*) of S_c in T_{i+1}. Then, we compare the size of the subtree c (n) with the size of the subtree rooted at the MRCA of S_c in T_{i+1} (m). If $m > \alpha n$ ($\alpha \in \{1.1, 1.2, 1.3, 1.5\}$) then we say that subtree c is successful (see Figure 5.3). Again we tried different α cutoffs to obtain a balanced dataset. We randomly choose 75% of the subtrees for training our model and leave the rest to test the model. In order to find the hyper-parameters of the model we performed 10 fold cross-validation on the training set. We tried *linear*, *radial* and *polynomial* kernels for the SVM and, among these, the linear kernel resulted in the best performance (see Figure 5.7).

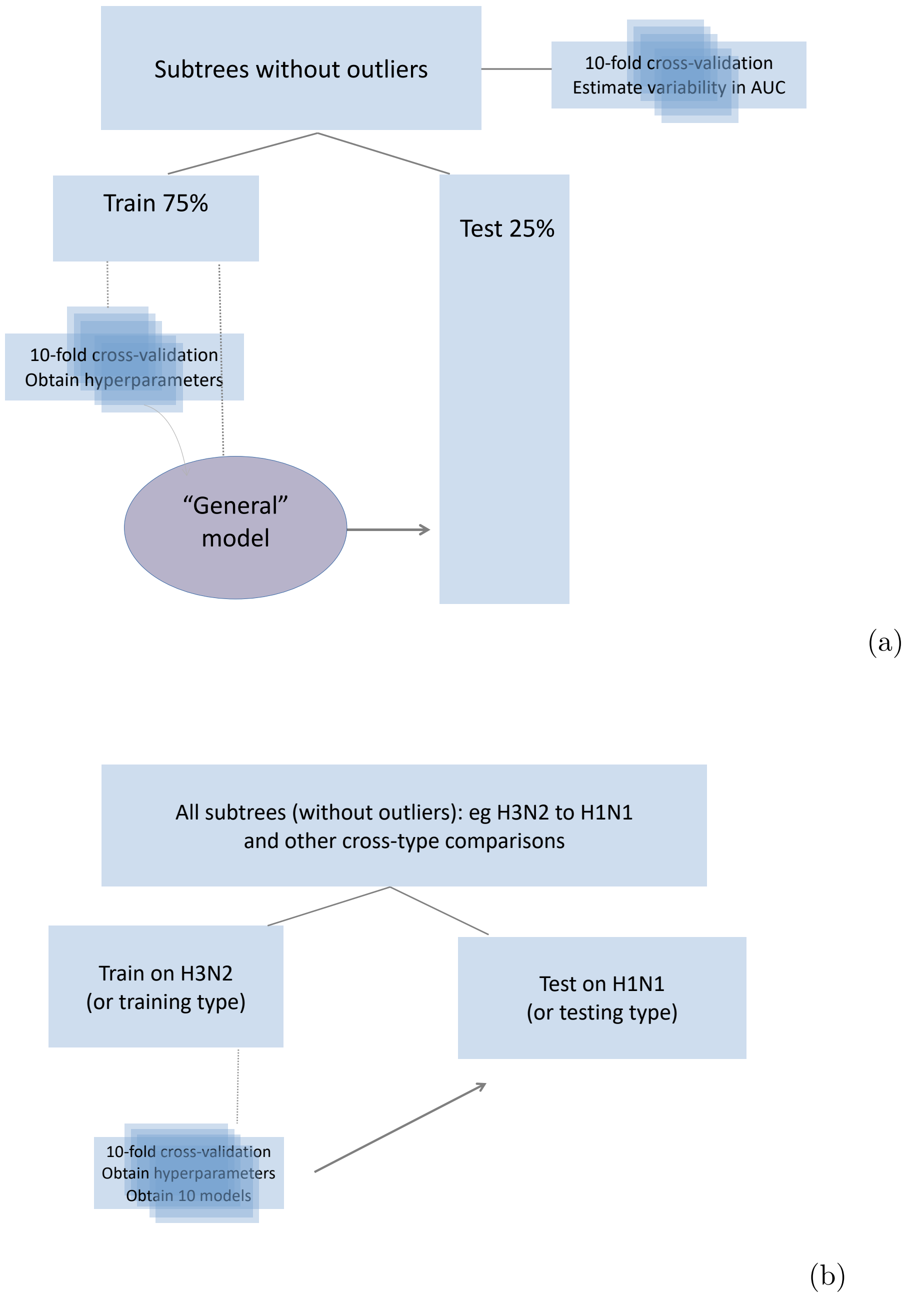


Figure 5.2: This figure illustrates our training and testing approach for experiments on influenza virus H3N2, influenza virus H1N1, influenza virus B and pooling the subtreees . We divide the subtrees whose outcome is known into training (75%) and testing (25%) data and choose hyperparameters using 10-fold cross-validation on the training data only. We use those hyperparameters to train the “general” model and test it on the testing data. (b) The schematic figure for training on a tree (H3N2) and testing on another tree (H1N1).

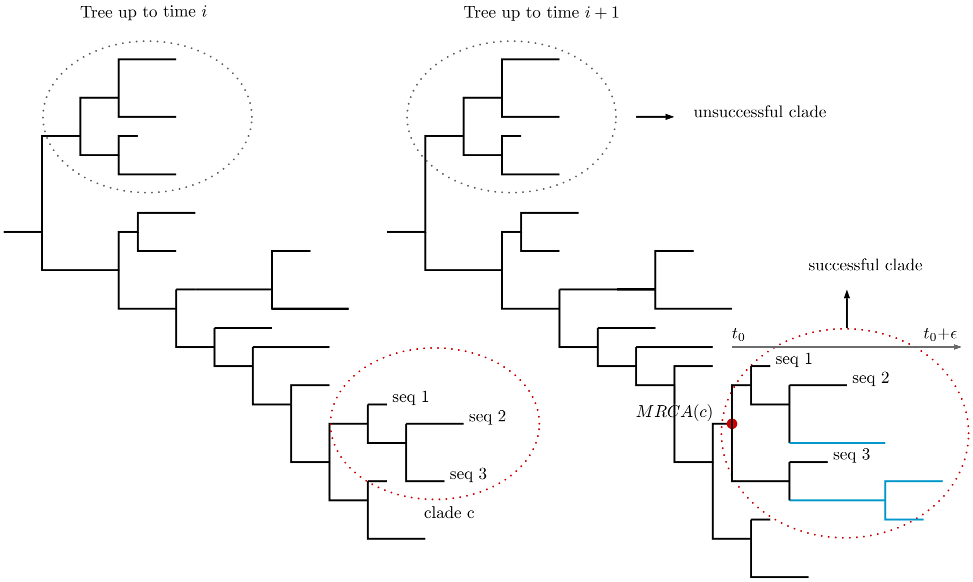


Figure 5.3: This figure depicts the definition of a successful subtree in the time slicing approach. The tree on the left, T_i, represents a tree reconstructed from a set of sequences up to time i and the tree on the right shows the same tree after one year (T_{i+1}). We compare these two trees to predict the successful subtrees in the time slicing approach. For each subtree c we compare the size of the subtree in T_i (X) with the size of the subtree rooted at the most recent common ancestor of the tips of subtree c in tree T_{i+1} (Y), in an interval of 3.4 years following the root of the subtree. If $Y > \alpha X$ ($\alpha \in \{1.1, 1.2, 1.3, 1.5\}$) then we say subtree c is successful. This overcomes the challenge that T_{i+1} does not contain the root of c; it does contain a node that is the MRCA of the tips in subtree c.

5.3 Results

Briefly, we extract subtrees from the H3N2, H1N1 and B phylogeny. Each subtree corresponds to an internal node of the tree and the tip descendants that have occurred within a fixed time frame (1.4 years). The remaining tips occur after the fixed time frame following the relevant internal node, and help to define whether the subtree has successfully grown into the future.

The approach results in a total of 396 subtrees, overlapping to some extent, containing 7615 of the 12785 tips in the full phylogeny. We use a wide range of features of the subtrees, focusing largely on tree structure but also including some branch length features, and the number of changes in the epitope sites of HA compared to previous sequences. We train supervised machine learning models to use this information to predict whether subtrees will succeed. Figure 5.4 shows the H3N2 hemaglutinin phylogeny and highlights in yellow the tips that belong to at least one subtree.

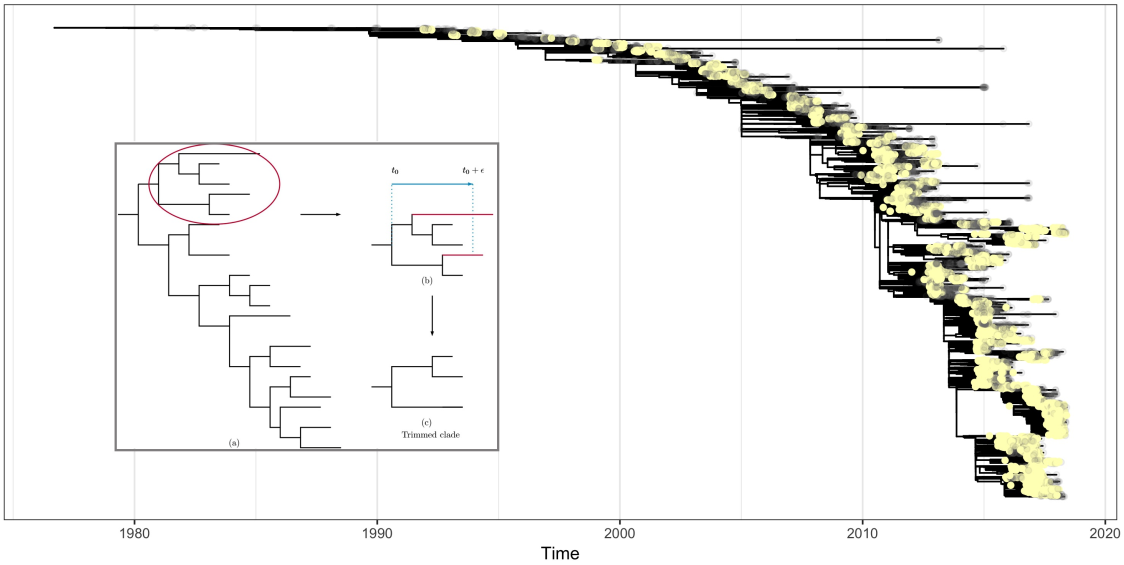


Figure 5.4: Phylogenetic tree reconstructed from H3N2 subtype sequences using RAxML, with tips highlighted. Each yellow tip is in a trimmed subtree (7615 out of 12785 tips); grey tips are not. The sequences are downloaded from GenBank with dates from 1980 to 2018-5. Long branches in this timed tree did not appear as long branches in the RAxML tree and were not removed (though their tips are not in any trimmed subtree). Inset: illustration of formation of trimmed subtrees: (a) the circled clade contains a subtree; (b) red branches reach tips that occur after the trimming time period and so are pruned out; (c) the resulting trimmed subtree.

The trained models successfully predict which subtrees will grow sufficiently, as measured within 3.4 years of a subtree's originating node (2 years after the last possible tip in a subtree). Using support vector machine (SVM) classification with a linear kernel, our overall 10-fold cross validation accuracy in H3N2 (using the HA sequences) was 74%. As the accuracy of a classifier can be misleading when there are uneven numbers of samples with the different outcomes, we use the area under the receiver-operator characteristic curve (AUC) to describe the overall performance of our models. The SMV had an average AUC of 0.82 (range 0.73-0.9); see Figure 5.6. We found an accuracy (portion correctly classified) of 79% and AUC of 0.89 when training on 75% of the subtrees chosen uniformly at random, and testing on the rest (Figure 5.5).

Figure 5.5 shows receiver-operator characteristic curves illustrating the trade-off between sensitivity and specificity. AUC ranges were obtained by training 10 models each on 90% of the subtrees; see Figure 5.6. We obtained a 79% accuracy and 0.86 AUC when we trained a linear kernel SVM model on a training portion of the subtrees (75% of the subtrees chosen uniformly at random) obtained from the H1N1 phylogeny, reconstructed using sequences from 2009 to 2018-05 (we did not use epitope features in any H1N1 analyses as the HA protein differs in H1N1). We performed 10-fold cross-validation on H1N1 subtrees, which

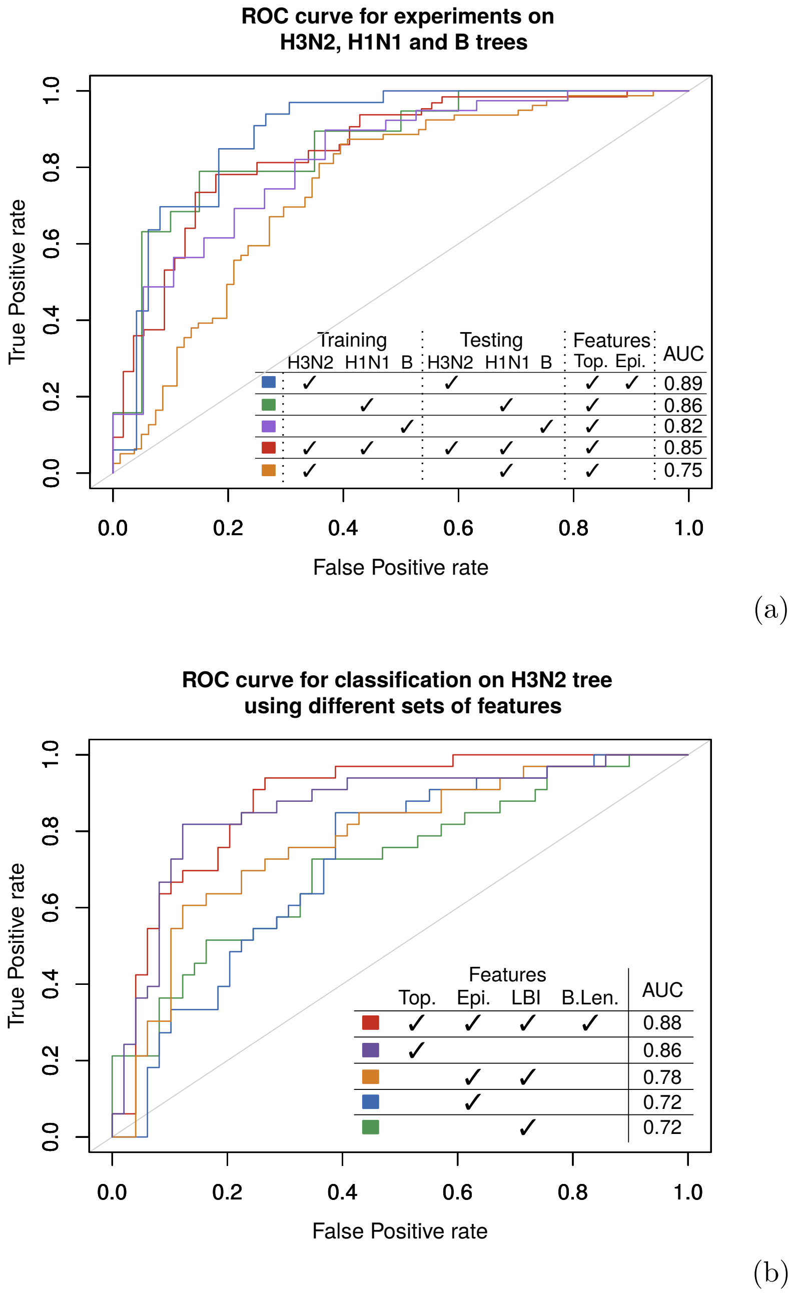


Figure 5.5: (a) ROC showing performance of linear kernel *SVM* trained and tested on H3N2, trained and tested on H1N1, trained and tested on B, trained and tested on the merged subtrees of H3N2 and H1N1 and trained on H3N2 and tested on H1N1. Figure 5.6 shows variation in these curves and their AUCs over 10 models trained on 90% of the test data for each case. (b) ROC showing classification performance for restricted sets of features. Top: tree shape features (no branch lengths). Epi: epitope features. LBI: local branching index. B.Len: some features include tree branch lengths.

resulted in 0.76 average AUC (range 0.52-0.95). We also pooled the subtrees of H1N1 and H3N2 and divided the pooled subtrees into training and test data sets; this resulted in an accuracy of 75% and an AUC of 0.85 (10-fold cross-validation AUC range 0.75-0.88). We also applied our method on a phylogenetic tree reconstructed from the HA gene of human influenza virus B. This results in a 0.82 AUC and 78% accuracy when we train our model on 75% of the data chosen randomly and test on the rest. For the influenza virus B tree, we used only the topological features. We also compared classifier performance using only portions of the data, and find that combining the tree shape features with epitope and local branching index (LBI) [131] gives the highest quality, with AUC of 0.88 compared to 0.72 for either epitope features [112] or LBI alone, and 0.78 for these combined.

Our subtrees are based on internal nodes in long-time phylogenies, and these nodes are present as a consequence of the relatedness patterns in all the data that are passed in to the tree reconstruction algorithm (in particular, including sequences from the entire time range). In consequence, a node's existence and local structure may be conditional on sequences occurring chronologically long after the node. We took several approaches to ensure that our models were not influenced by some such subtle knowledge of the future. We trained models on H3N2 HA phylogeny but tested on an H1N1 HA phylogeny (Figure 5.2(b)). We obtained an accuracy of 72% and an AUC of 0.75 (range 0.72-0.75 when training 10 models each on 90% of H3N2 subtrees and testing on H1N1; Figures 5.5 and 5.6). The reduced accuracy is natural given that the HA proteins differ between the two types. We also created "time slices" from the H3N2 HA sequences using only tips occurring prior to time i ($i \in \{2012-5, 2013-5, 2014-5, 2015-5, 2016-5, 2017-5\}$). We extracted subtrees, and tested their success using trees reconstructed from all the sequences prior to time $i+1$ respectively (Figure 5.3). This mimics a "real-time" analysis and ensures that subtrees cannot depend on sequences arising after a set time. This approach performs comparably to our other tests, with accuracy of $70\% - 76\%$ and an AUC of $0.73 - 0.86$ (Figure 5.7).

5.3.1 Cross-Validation on Different Experiments

To determine how variable the AUC and accuracy figures are for our analyses, we performed 10-fold cross validation for the main prediction on H3N2, the analysis on H1N1 only (trained on a portion of the H1N1 subtrees), the pooled analysis of H3N2 and H1N1 and the predictions on H1N1 using a model trained on H3N2. The resulting ROC curves are shown in Figure 5.6(a-d) respectively. AUC values are consistently above the random classifier's expected value of 0.5, with ranges given in the caption of Figure 5.6. It is particularly encouraging that the test on H1N1, using a model trained on H3N2 (Figure 5.6(d)) has a highly consistent set of AUC and accuracy values (range of AUC 0.72-0.75) because this test is in some sense the most challenging; no subtree in the test data (H1N1) ever shares a tip with a subtree in the training data (all H3N2) and the risk of overfitting is low (also note that we did not select a model or method based on this test). The most variable set of AUCs arises from H1N1 alone, which is likely due to the lower volume of data.

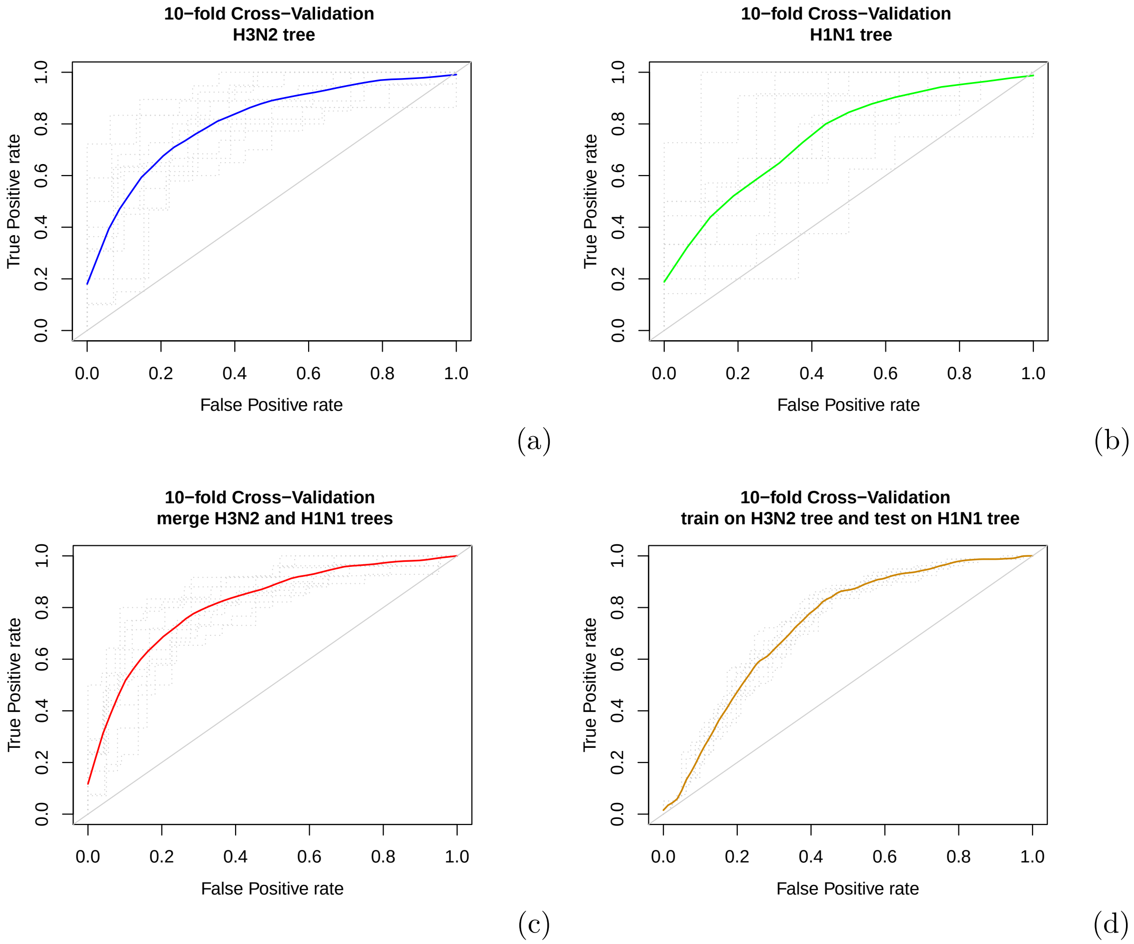


Figure 5.6: (a) The result of 10-fold cross validation for *SVM* models trained on the subtrees whose ancestral nodes occurred pre-2015-1 or which began later but whose outcomes are known. AUC values range from 0.73 to 0.90 (average 0.82) with 80% of the folds resulting in more than 0.75 AUC. (b) 10-fold cross-validation on the H1N1 tree using *SVM* with linear kernel and a set of topological properties of the clades. AUC values range from 0.52 to 0.95 (average 0.76) with AUC more than 0.75 in 70% of the folds. (c) The result of 10-fold cross-validation on the merged dataset of H3N2 and H1N1 trees. The minimum and maximum values of AUCs among the folds are 0.75 and 0.88 respectively (average 0.82). (d) 10-fold cross-validation for training on the H3N2 tree and testing on the H1N1 tree; we removed 10% of the training data at each fold and evaluated the model on the test data. AUC values range from 0.72 to 0.75 (average 0.74).

5.3.2 Results of Time Slices

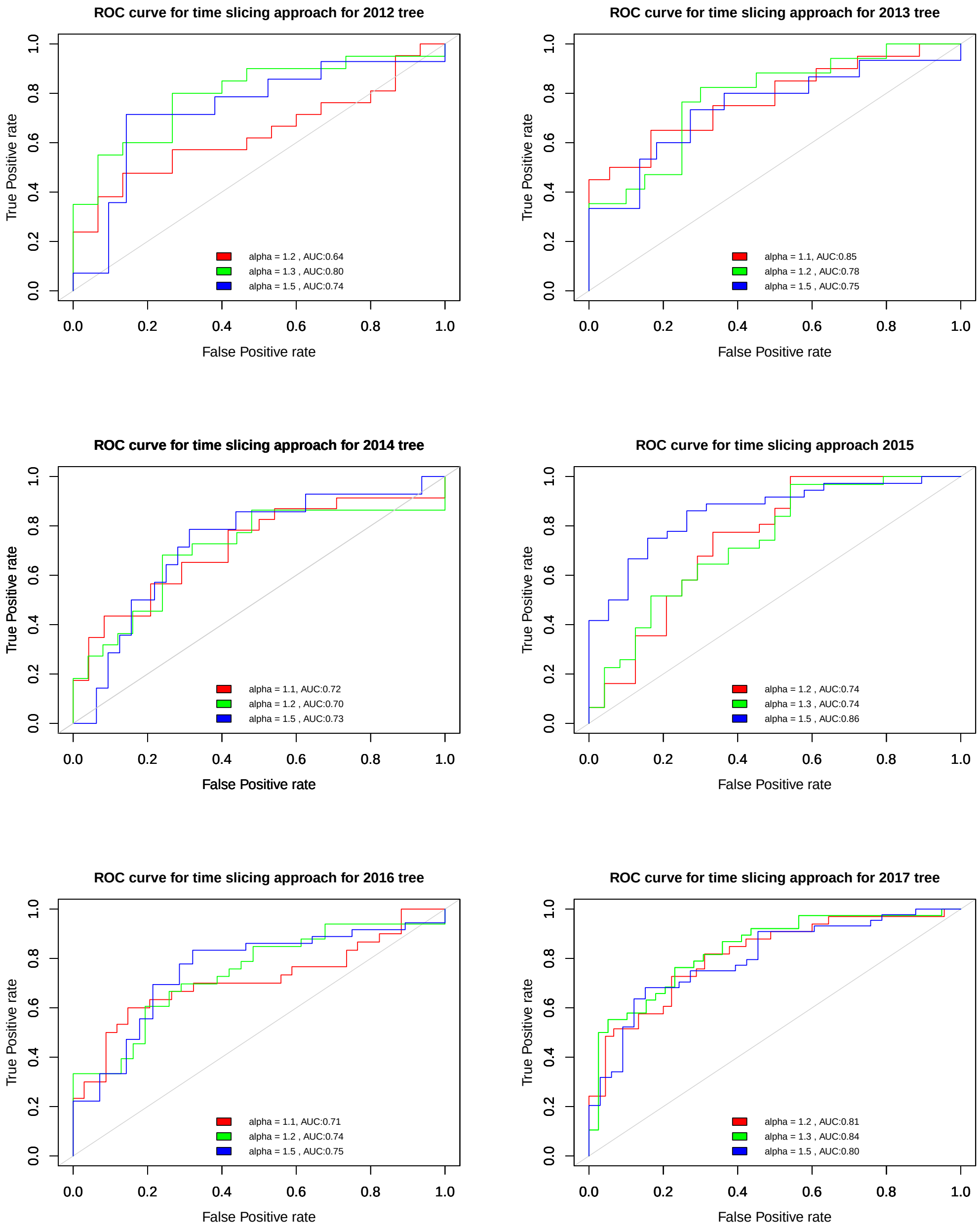


Figure 5.7: We classified the subtrees of the influenza virus H3N2 tree reconstructed from the sequences up to time i by comparing with the tree reconstructed from sequences up to time $i+1$ which $i \in \{2012-5, 2013-5, 2014-5, 2015-5, 2016-5, 2017-5\}$. In this figure the result of classification using different ratio between the size of a subtree in time i and the size of the corresponding subtree in time $i+1$ is shown.

5.3.3 Model Parameters

Because we have used a fixed time frame to select subtrees, the subtrees vary considerably in size (since we do not control size). Successful subtrees are larger (median 18) than unsuccessful ones (median 12). Figure 5.8 shows the sizes of the trimmed subtrees from H3N2 that were used in the main analysis, with bars shaded to the outcome. We observed that size alone does not successfully classify the success of subtrees, though it contains some of the information (AUC 0.63). We chose to control time frame rather than subtree size, as time frame has a clear biological meaning and the size of subtrees may in fact be useful information for the classification; trees with rapid branching can achieve more tips in the fixed time frame. However, size (and apparent rapid branching) may also reflect sampling differences.

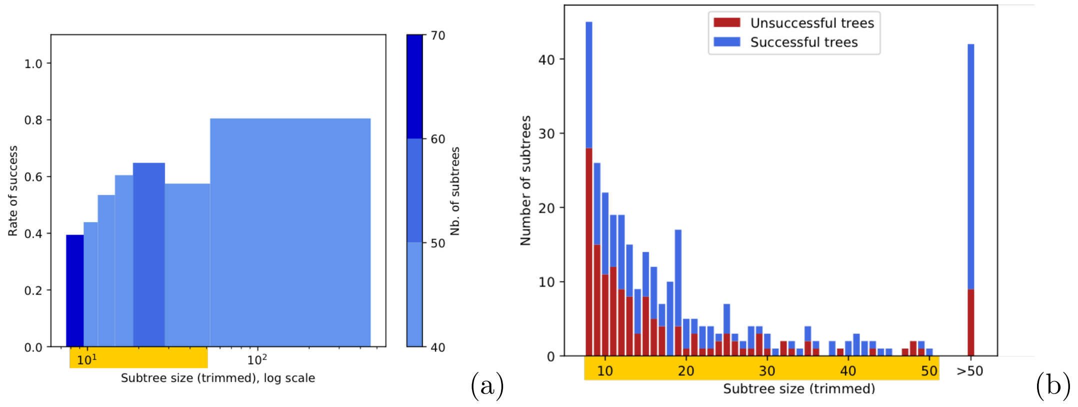


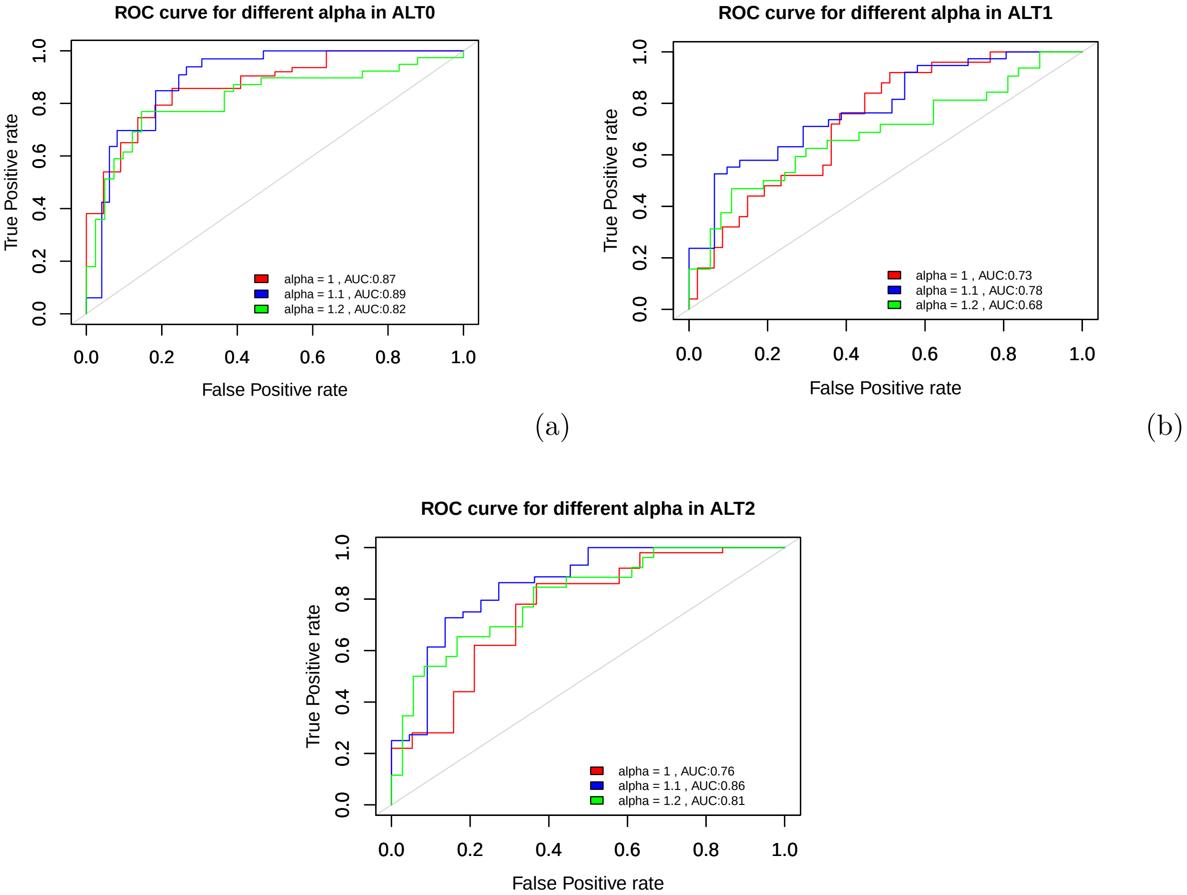
Figure 5.8: (a) Correlation between the size of trimmed subtrees (x-axis) and the rate of success (y-axis) for the 391 subtrees from the H3N2-HA dataset. The rate of success is defined as the number of succesful trees divided by the total number of trees for a certain range of sizes. The ranges were computed in order to encompass approximately the same number of subtrees, and the color of the bars represent how many subtrees were taken into account for the computation of the success rate. The subtrees vary from size 8 up to 460, but as most of the dataset is composed by small subtrees, we used a log scale to better visualize the information. The subtree sizes highlighted in yellow are detailed in the bottom panel. (b) Subtree size distribution for trees up to size 50, which corresponds to 87.4% of the dataset. For each size, the graph shows how many subtrees were succesful (blue blar) and unsuccesful (red bar).

We used three sets of parameters to extract the subtrees of influenza virus H3N2. These three models are summarized in Table 5.2, and the results of prediction using each of these models are shown in Figure 5.9.

We also explored changing the success threshold, defining a subtree of size n as successful if its ancestor eventually has at least one more tip, more than $1.1n$ tips and more than $1.2n$ tips (ie if size $> \alpha n$ for $\alpha = 1, 1.1, 1.2$). Figure 5.9 shows the performance under these variations.

Model	MinTotalSize	MinTrimSize	OverlapCutoff	TimeFrame
ALT0	8	8	0.8	1.4
ALT1	12	12	0.95	2
ALT2	7	7	0.7	1

Table 5.2: Our subtree selection algorithm uses three parameters: a minimum subtree size, a maximum allowable overlap and the length of the time window (Figure 6.1). We denote our default as “ALT0” and our alternatives as “ALT1” and “ALT2”. There are natural trade-offs: a larger minimum size, lower overlap and longer time frame all result in fewer accepted subtrees. We found good performance for each of these three alternate setups.



(a)

(b)

(c)

Figure 5.9: The ROC curves from prediction for the H3N2-HA dataset using different models and different definitions of a successful subtree. In all models, setting the threshold to an appropriate value allows a balance between successful and unsuccessful outcomes, resulting in better performance. Among different sets of parameters for extracting the subtrees, ALT0 is the most powerful model resulting in 0.89 AUC. In all of these experiments we used the topological and epitope features.

5.3.4 Influenza Virus B

We applied our method to influenza virus B. We used the same approach as for H3N2 and H1N1 to reconstruct the influenza virus B tree using sequences from 1980 to 2018-05 (for further details see Section 5.2.1 The results of different experiments on influenza virus B are shown in Figure 5.10.

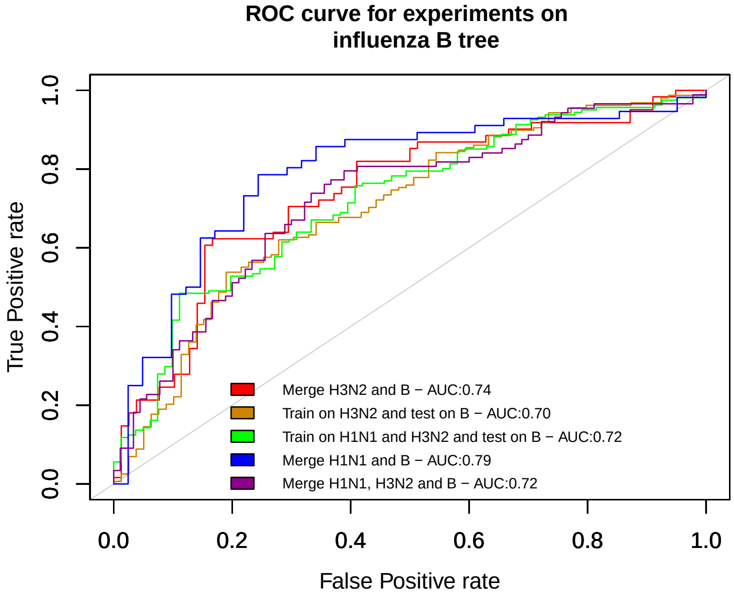


Figure 5.10: We find that a choice of $\alpha = 1.2$ results in a good balance between successful and unsuccessful subtrees. Combining H1N1 and B subtrees results in good performance, suggesting similarities between the link between subtree structure and success between H1N1 and influenza virus B.

5.3.5 Best Features

For robust feature selection we followed the ensemble technique introduced in [164]. They show that combining multiple (unstable) feature selectors yields more robust feature selection than using a single selection method. We use 4 models including logistic regression, random forests, SVM with linear kernel and learning vector quantization (LVQ)[95] to rank the features based on their contribution in the classification task (see Figure 5.11). In the general classification of the H3N2 subtrees, the epitope features are among the most important. However, classification based only on epitope features reduces the AUC from 0.89 to 0.72, and our classifiers perform well on H1N1 (0.86 AUC, and 0.76 average AUC in the 10-fold cross validation) despite not having the epitope features. No single feature or small group of features that we have identified can perform as well as the combined phylogenetic and epitope features. We did not attempt to reduce the feature set to obtain a minimal set of features with the optimal performance.

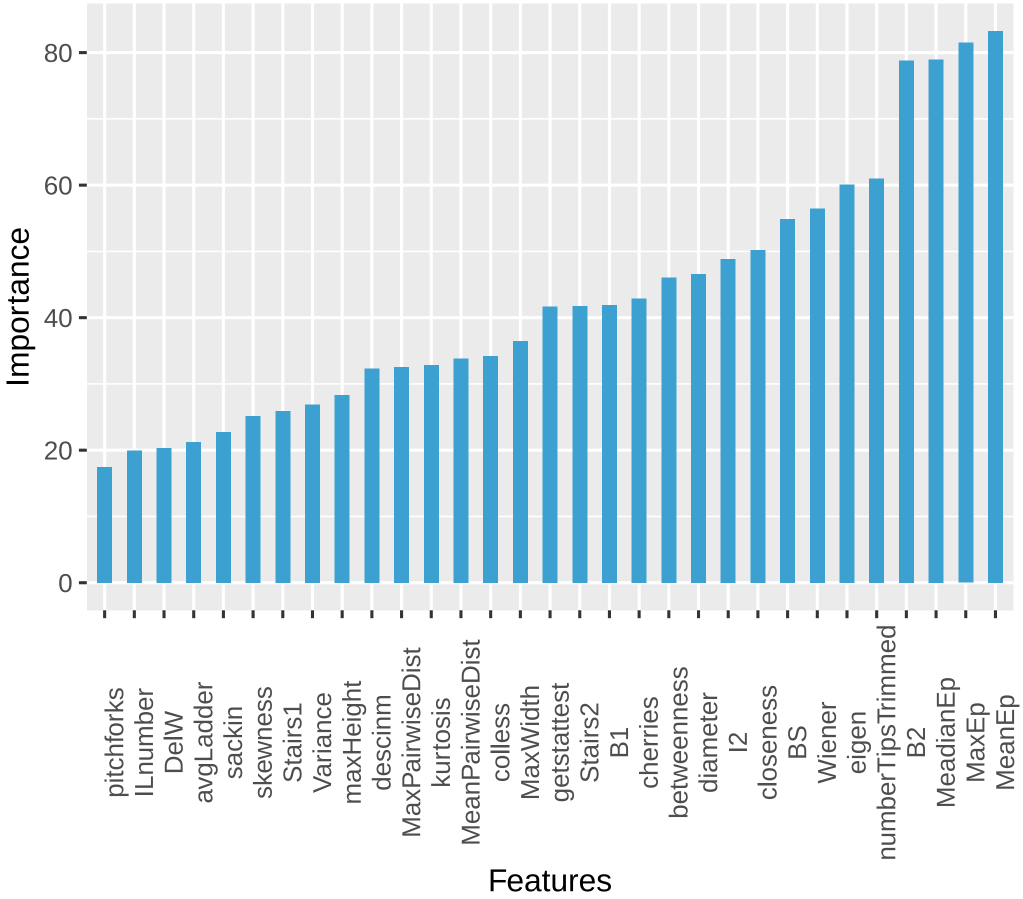


Figure 5.11: This figure shows the importance of each feature in the classification task. We use an ensemble technique to find the importance of each feature. MeanEp, MaxEp, and MeadianEp are the three features we defined on the epitope sites of the sequences (see Section 5.2.3 for more details). numberTipsTrimmed is the number of tips in the trimmed subtrees. The rest of the features are defined in the Table 5.1.

5.3.6 Predictions on the Recent Subtrees

Subtrees originating after January 2015 (here called “recent subtrees”) did not have enough time to grow into the future, and we do not know whether they are successful, as the predictions are relative to 3.4 years after the initial node. To accommodate for this, throughout our analysis, we only trained and tested models on subtrees that originated prior to January 2015. In this second part, our aim is to make predictions for subtrees whose outcomes are not known. In order to do so, we trained 10 models using 10-fold cross-validation on the non-recent subtrees, as well as a general model, and used these 11 models for predictions on the recent subtrees (Figure 5.12).

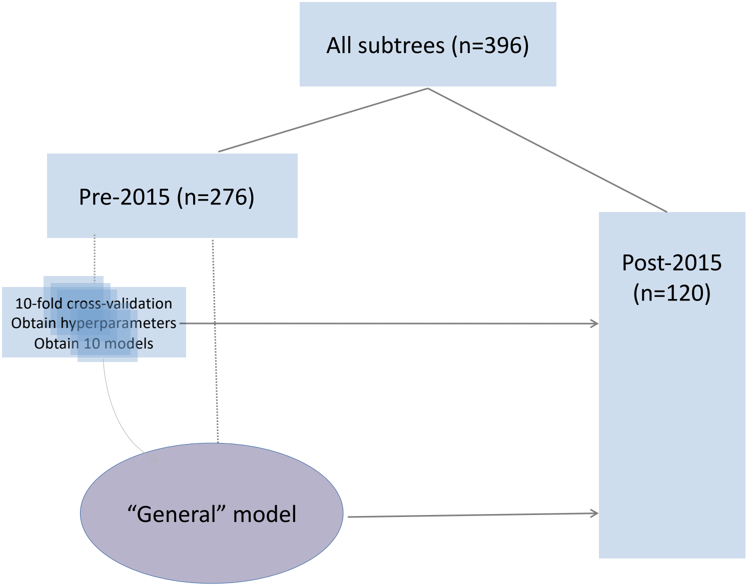


Figure 5.12: For the prediction task, we divide our subtrees into recent (subtrees originating after 2015-1) and non-recent sets (the rest of the subtrees). We then trained 10 models using 10-fold cross validation together with one general model on the set of non-recent subtrees and used these 11 models for prediction on the recent subtrees.

The recent subtrees were labeled according to the clades defined by the World Health Organisation (WHO) (see Figure 5.13(a)). Every recent subtree contributed with 11 predictions to its respective clade. The predicted success of a clade in 2019 is the average success of all predictions coming from related recent subtrees. The predictions are presented in figures 5.13 and 5.14.

Our recent subtrees originated between 2015-03 and 2017-02, and in this period, contained tips as shown in Figure 5.13(b), with the majority of tips in clades A1b, A1, A1a, A2 and A2/re. This reflects the sequences in GenBank, and is likely not globally representative [72]. Clades A3, A1 and A1b/135K and A2/re were most strongly predicted to be successful by our measure (fraction of the clade's subtrees predicted as successful), but in A1b/135K there are only two subtrees on which to base predictions. In clade A3 we predicted 7/8 subtrees as successful, with 5 of those already having shown sufficient growth to meet our success criterion. A2/re has an intermediate signal overall, but has 12 subtrees predicted to be successful. Of these, 10 had already shown sufficient growth to pass our success threshold by the end of our sampling. Indeed the A2/re clade did become very successful, probably due to a re-assortment event [11]. Our model also predicted that other parts of clade A2 (4 of 7 subtrees in A2 but not in A2/re) may grow. In the time frame we had, there were relatively few sequences in our GenBank data that were mapped into the A4, A1b/135K and A1b/135N clades by the 'augur' pipeline [72] so these clades have very few subtrees on which to make predictions.

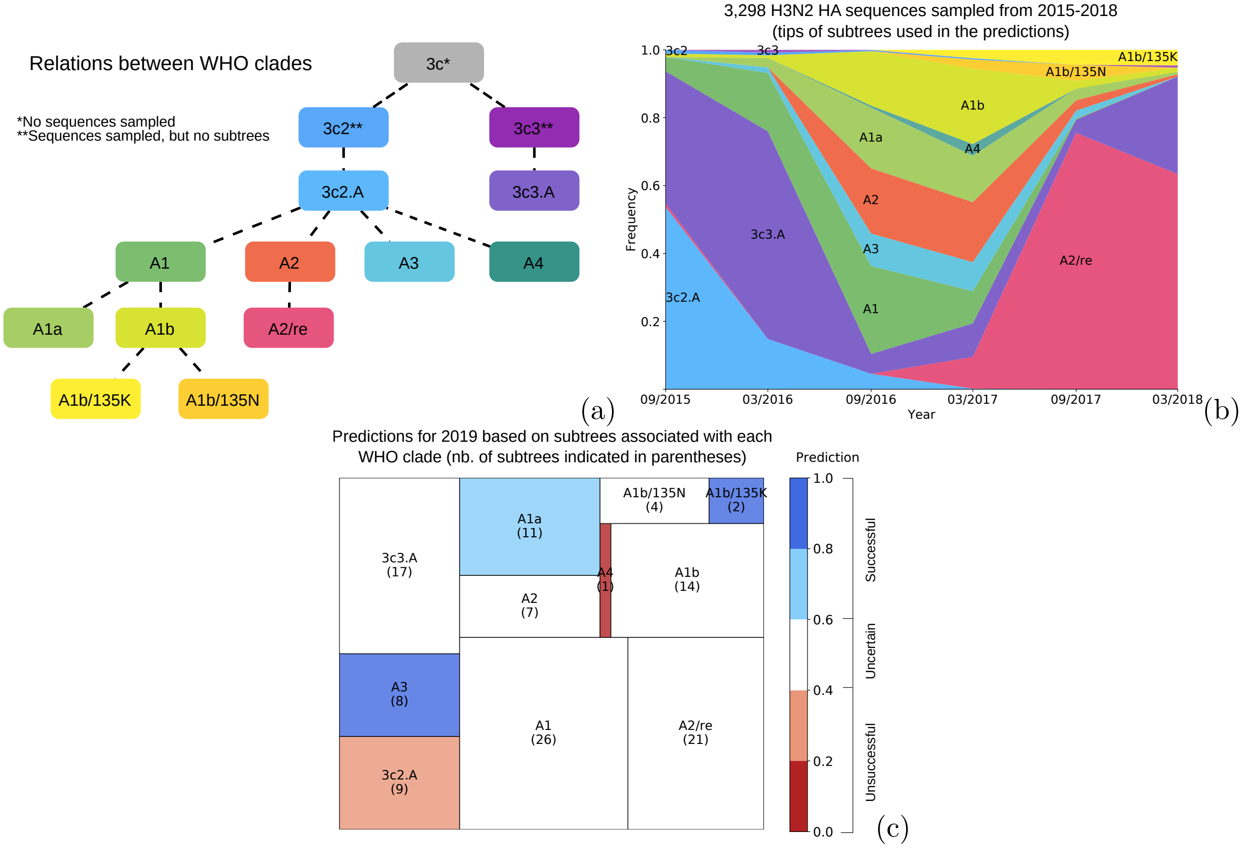


Figure 5.13: Summary of predictions for the recent subtrees (for detailed information, see Figure 5.14). (a) Relations between the clades defined by the World Health Organisation (WHO), showing the emergence of new clade designations from existing ones; (b) Frequency of clades through years, based on 3298 H3N2 HA sequences sampled between 2015 and 2018. Every sequence is associated with a tip of one or more subtrees used in the predictions. The clade designation of the tips determines the clade designation of the subtrees by majority rule. (c) Predictions for the recent subtrees. Among the 120 recent subtrees, the outcome (success/failure) of 63 subtrees is not known. Each subtree was tested on 11 different SVM models (see Figure 5.12). A rectangle corresponds to the prediction of a clade, and its area is proportional to the number of subtrees used in the prediction. The number of subtrees associated with a clade is indicated in parentheses. Color reflect the combined predictions of the subtrees associated with each clade.

			Clades										
			A1b/135K	A3	A1a	A2/re	A1b	A1b/135N	A2	3c3.A	A1	3c2.A	A4
Prediction for 2019 (rate of success)			1.000	0.875	0.653	0.593	0.591	0.545	0.493	0.487	0.465	0.232	0.182
Nb. of subtrees used in predictions			2	8	11	21	14	4	7	17	26	9	1
Details of subtrees used in predictions: Current growth	Prediction												
Below success threshold	Successful	A	1	2	5	2	2	1	4	1	4	0	0
	Unsuccessful	B	0	1	2	7	5	2	3	5	11	5	0
Above success threshold	Successful	C	1	5	3	10	6	1	0	7	8	2	0
	Unsuccessful	D	0	0	1	2	1	0	0	4	3	2	1

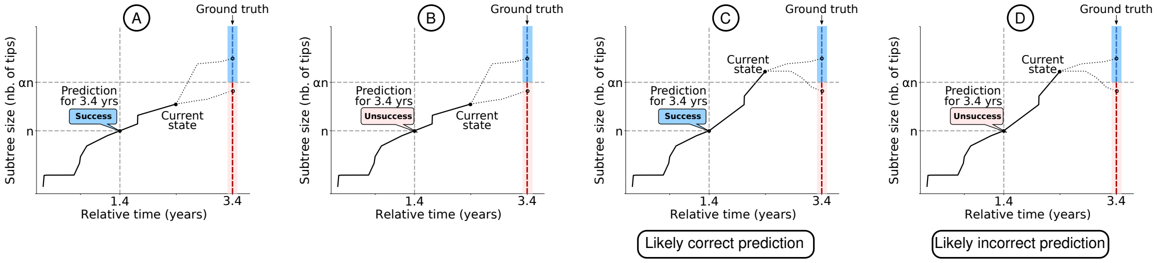


Figure 5.14: Predictions of the rate of success for the clades defined by the World Health Organisation (WHO). A column contains the information about one clade, indicated on the top and following the same color code as Figure 5.13. The columns (clades) are sorted according to the rate of success, from the most successful (left) to the least successful (right). The rate of success of a clade is an average of all predictions from subtrees whose majority of tips contain DNA markers linked to the clade (11 predictions per tree). All subtrees used in these predictions were extracted between 2015 and 2018 and, thus, have an unknown outcome, as the actual growth rate in 3.4 years from the initial node is still inaccessible. The subtrees are classified in four types: (A) subtrees whose prediction is to succeed but the current growth is still below the threshold considered to be successful; (B) subtrees whose prediction is to fail and the current growth is below the threshold considered to be successful; (C) subtrees whose prediction is to succeed and the current growth is above the threshold considered to be successful (notice that even in this scenario it is not possible to be certain, since the structure of the tree may change); (D) subtrees whose prediction is to fail but the current growth is above the threshold considered to be successful. Schematics for each case are presented on the bottom of the table.

Chapter 6

Computing the Distribution of the Robinson-Foulds Distance

One of the crucial aspects of phylogenetics is the comparison of two or more phylogenetic trees. There are different metrics for computing the dissimilarity between a pair of trees. The Robinson-Foulds (RF) distance is one of the widely used metrics on the space of labeled phylogenetic trees. We do not claim that the RF distance is the most practical one in phylogenetics; indeed this distance may be highly biased [109]. Instead, the aim of this chapter is to show one of the applications of the DFT in improving the running time of an existing algorithm. Pattengale et al. [142] introduced a novel randomized approximation algorithm for computing a high probability approximation of the true RF distance in sublinear time. However, the best known algorithm for computing its distribution - i.e. the frequency of each possible value over the space of all possible trees with n tips - requires $O(n^5)$ time, which becomes prohibitively slow even for moderate n.

This distribution arises naturally when one is interested in performing statistical tests. For instance, one could easily use it to compute the probability of the distance between two given trees being larger than the distance between the first tree and a tree chosen uniformly at random (a null model). In other words, this distribution informs us about how likely the observed distance between two trees is to occur by chance [26, 184]. In [76, 182], the authors propose a method for computing the distribution of the RF distance between a given tree T_0 and all the trees on the same number of tips and having the same labels, using generating functions. The approach does not result in a polynomial-time algorithm [26]. Bryant and Steel, whose work serves as the basis of our approach, have proposed a polynomial-time algorithm via a dynamic programming approach for computing the distribution of the RF distance from a given T_0 [26]. They also showed that a Poisson distribution whose parameter depends on the number of cherries of T_0 can approximate it well when n is large.

Although their algorithm runs in polynomial time, it is quintic in the number of tips. However, the main bottleneck of their dynamic programming formulation can be expressed as a matrix convolution [49]. This immediately suggests the use of a fast convolution computation, such as the Discrete Fourier Transform (DFT) or the Fast Fourier Transform (FFT) [36], to reduce the running time to essentially cubic, and make the algorithm practical. Unfortunately, the FFT approach suffers from significant numerical stability issues, and often produces nonsensical negative values. We thus turn to an alternative, called the Number-Theoretic transform (NTT) [196], which is the generalization of the FFT to finite fields. This allows us to keep the running time essentially cubic while completely eliminating all numerical stability issues. From now on, by DFT we, mean the Discrete Fourier Transform over the field of complex numbers and by NTT we mean the Number-Theoretic transform, i.e. the Discrete Fourier Transform over the number field F_p formed by the integers modulo the prime p.

The DFT has been used in computational biology in the past, to solve problems such as locally aligning two sequences [127], computing the statistics (such as p-values) of a pairwise local alignment [55] or multiple alignment [155], as well as solving problems in mass spectrometry [165] (an area where, incidentally, the term "spectral convolution" has a different meaning [171]). In these applications, the main use of the DFT is to quickly compute convolutions between long vectors. However, there has been, to our knowledge, no computational biology paper using the Number-Theoretic Transform (NTT) since 1990 [14]. We believe that this is a missed opportunity. Our hope is that the reintroduction of the NTT in the context of phylogenetics will inspire other computational biology researchers to apply it to solving their problems.

6.1 Problem Definition

The distribution of the RF distance from a given unrooted labeled phylogenetic tree has been studied before, but the fastest known algorithm for computing this distribution is a slow, albeit polynomial-time, $O(n^5)$ algorithm. We modify the dynamic programming algorithm for computing the distribution of this distance for a given tree by leveraging the Number-Theoretic Transform (NTT), and improve the running time from $O(n^5)$ to $O(n^3 \log(n))$, where n is the number of tips of the tree. In this chapter, we only consider labeled phylogenetic trees.

6.2 Methods

We start with a detailed explanation of the method proposed by Bryant and Steel [26] to compute the distribution of the RF distance from a fixed tree T to all the trees on the same set of labels. We then introduce our approach for improving the running time of their algorithm, which makes it practical for moderate to large size trees.

Given a phylogenetic tree T, we distinguish between three types of internal nodes in a rooted binary tree. Type I nodes are internal nodes with two tip descendants (also called *cherries*), type II nodes are internal nodes with one tip and one internal node as descendants, and type III nodes are internal nodes with two internal node descendants (Figure 6.1). The expected number of cherries (type I internal nodes) in a binary rooted tree with n tips under the equal rate Markov model *(ERM or Yule)* [73, 204] is $n/3$ [118]. The expected number of the other types are given in Lemma 2.

Lemma 2. *The expected numbers of type II and type III internal nodes under the Yule model are* $\frac{n}{3}$ *and* $\frac{n}{3} - 1$*, respectively.*

Proof. Let N_1, N_2 and N_3 denote the expected number of cherries, type II and type III internal nodes of the tree under the *Yule* model respectively. We have $2N_1 + N_2 = n$ and since $N_1 = \frac{n}{3}$ then we have $N_2 = \frac{n}{3}$. Considering the fact that the number of internal nodes in a binary rooted phylogenetic tree with n tips is $n-1$, we can show that $N_3 = n-1-2\frac{n}{3} = \frac{n}{3} - 1$. □

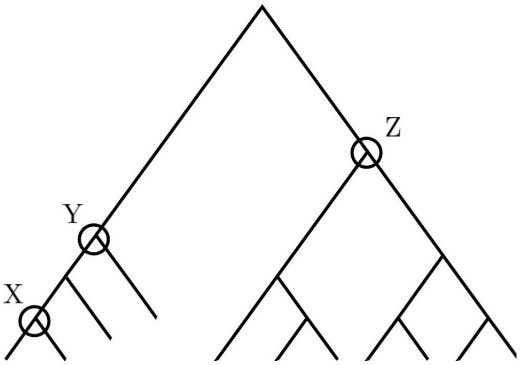


Figure 6.1: There are three types of internal nodes in a binary rooted tree. X is a type I node, i.e. an internal node with two tip descendants (cherry). Y is a type II internal node since it has one tip and one internal node descendants. Z is a type III internal node since it has two internal node descendants.

6.2.1 Computing the Distribution of the RF Distance

Our method to compute the distribution of the *RF* distance is based on the simplification of the dynamic programming approach introduced by Bryant et al. [26]. We use the same notation and definitions that they do.

For a given unrooted labeled tree T with n tips, let us denote by $b_m(T)$ the number of unrooted labeled phylogenetic trees which are at a distance m from T. The generating function of $b_m(T)$ can be computed by the following recursive formula [76]:

$$B(T, x) = xB(T/e, x) + (1 - x^2)B(T_1, x)B(T_2, x),$$

where $B(T,x) := \sum_{m\geq 0} b_m(T)x^m$. Here, e can be any internal edge, T/e represents the tree after contracting e, and T_1 and T_2 are the trees obtained form T after removal of e. The exponential number of distinct sub-cases in the above recursion precludes a polynomial-time algorithm to compute the distribution of the RF distance directly based on this recursion [26, 76].

An internal split is a split which is obtained by removing an internal edge from T. By $q_s(T)$, we denote the number of unrooted labeled phylogenetic trees with n tips which have exactly s internal splits in common with T. Bryant and Steel [26] showed that

$$b_m(T) = q_{(n-3-m/2)}(T),$$

where $m = 0, 1, \ldots, 2(n-3)$. Let us define the polynomial with coefficients $q_s(T)$:

$$q(T,x) = \sum_{s=0}^{n-3} q_s(T)x^s.$$

Consider a subset of internal edges $E \in \mathcal{E}(T)$, and suppose that removing E from T results in $|E|+1$ connected components $T_1, T_2, ...T_{|E|+1}$. By $\mathcal{E}(T_i)$, we denote the number of internal edges of T that are contained in T_i. Define

$$N_E(T) = \prod_{i=1}^{|E|+1} \mathcal{B}(|\mathcal{E}(T_i)|),$$

where $\mathcal{B}(m)$ is the number of unrooted labeled phylogenetic trees with m internal edges. From this definition, it is clear that $N_E(T)$ is equal to the number of unrooted labeled phylogenetic trees that contain the splits induced by the edges in E [26]. It can be easily seen [170] that $\mathcal{B}(m) = b(m+3) = \prod_{k=3}^{m+3}(2k-5)$, where $b(n)$ is the number of unrooted labeled binary trees with n tips.

$$b(n) = (2n-5)!! = \prod_{k=3}^{n}(2k-5).$$

Finally, for $s \geq 0$, we denote by $r_s(T)$ the sum of $N_E(T)$ over all possible subset of edges E with $|E| = s$:

$$r_s(T) = \sum_{E\in\mathcal{E}(T), |E|=s} N_E(T).$$

Bryant et al. [26] uses a dynamic programming approach to compute the distribution of the RF distances from a given unrooted labeled phylogenetic tree T with n tips as the first argument. Denote the node adjacent to tip n in T by v_0. Remove tip n, and root the resulting tree with v_0 as the root. We use this rooted labeled tree as the input to the dynamic programming algorithm. From now on by T, we mean this new rooted labeled tree, and by T_v we denote the subtree rooted at an internal node v of T.

For $s, k \geq 0$, let $\mathcal{A}(v, s, k)$ be the set of all subsets $E \in \mathcal{E}(T_v)$ such that $|E| = s$ and the number of internal edges of T in the component of $T_v - E$ containing v equals k. Define

$$R(v, s, k) := \sum_{E \in \mathcal{A}(v,s,k)} N_E(T_v).$$

Figure 6.2 illustrates the computation of $R(v, s, k)$. Note the assumption that T is fully resolved, and before counting the internal edges of each component, all internal nodes of degree 2 are removed (since removing a set of edges from T might result in some components with internal nodes of degree 2).

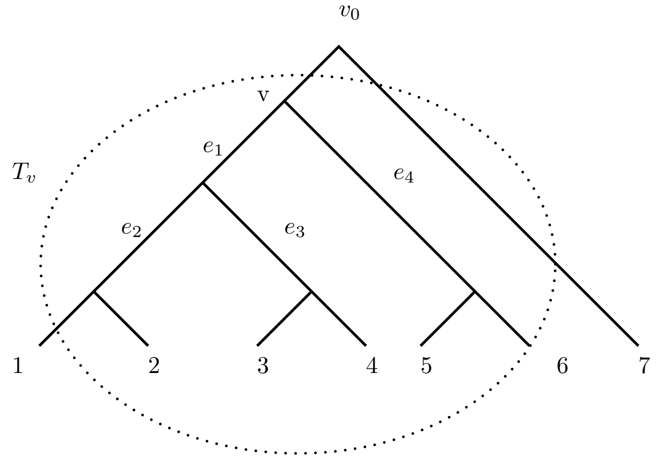


Figure 6.2: This figure is an example of how to compute $R(v, s, k)$. Consider the subtree rooted at v. For $k = 1$ and $s = 2$ there are four options for E: $E_1 = \{e_2, e_4\}$, $E_2 = \{e_3, e_4\}$, $E_3 = \{e_1, e_2\}$, and $E_4 = \{e_1, e_3\}$. The value of $N_E(T_v)$ for each of these sets is $\mathcal{B}(0)\mathcal{B}(1)\mathcal{B}(0)$, $\mathcal{B}(1)\mathcal{B}(0)\mathcal{B}(0)$, $\mathcal{B}(0)\mathcal{B}(1)\mathcal{B}(1)$ and $\mathcal{B}(1)\mathcal{B}(0)\mathcal{B}(1)$ respectively, so $R(v, s, k) = 24$

For $v = v_0$, we have the following equation for $r_s(T)$:

$$r_s(T) = \sum_{k=0}^{n-3-s} R(v_0, s, k).$$

We mention in passing that the upper bound for this summation was incorrectly given as s by Bryant and Steel. The correct bound is $n - 3 - s$ since after s internal edges are removed from T, the component containing v_0 contains at most $n - 3 - s$ internal edges.

From [182] we have the following equality for the generation function of $r_s(T)$:

$$R(T, x) := \sum_{s \geqslant 0} r_s(T) x^s \implies q(T, x) = R(T, x - 1). \tag{6.1}$$

Using these definitions, Bryant et al. developed a recursion for computing $R(v, s, k)$. For every $v \in \mathcal{I}(T)$, $s, k \geq 0$, this recursion is derived through the following lemmas:

Lemma 3. *[26]*

$$R(v,0,k) = \begin{cases} \mathcal{B}(k) & \text{if } k = \mathcal{E}(T_v), \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

Lemma 4. *[26]*

1. *$R(v, s, k) = 0$, if $k > |\mathcal{E}(T_v)|$ or v has no children in $\mathcal{I}(T)$ and $s \geq 1$*
2. *If v has one child in $\mathcal{I}(T)$ say v_1*

$$R(v,s,k) = \begin{cases} \sum_{k_1 \geq 0} R(v_1, s-1, k_1) & if k = 0, \\ R(v_1, s, k-1)(2k+1) & \text{otherwise.} \end{cases} \tag{6.3}$$

3. *If v has two children in $\mathcal{I}(T)$ say v_1 and v_2 and $k = 0$*

$$R(v,s,0) = \sum_{s_1=0}^{s-2} (\sum_{k_1 \geq 0} R(v_1, s_1, k_1))(\sum_{k_2 \geq 0} R(v_2, s-2-s_1, k_2)). \tag{6.4}$$

4. *If v has two children in $\mathcal{I}(T)$ say v_1 and v_2 and $k \geqslant 1$*

$$\begin{aligned} R(v,s,k) = \sum_{s_1=0}^{s-1} (\sum_{k_1 \geq 0} R(v_1, s_1, k_1)) R(v_2, s-1-s_1, k-1) \frac{\mathcal{B}(k)}{\mathcal{B}(k-1)} \\ + \sum_{s_2=0}^{s-1} (\sum_{k_2 \geq 0} R(v_2, s_2, k_2)) R(v_1, s-1-s_2, k-1) \frac{\mathcal{B}(k)}{\mathcal{B}(k-1)} \\ + \sum_{s_1=0}^{s} \sum_{k_1=0}^{k-2} R(v_1, s_1, k_1) R(v_2, s-s_1, k-2-k_1) \frac{\mathcal{B}(k)}{\mathcal{B}(k_1)\mathcal{B}(k-2-k_1)}. \end{aligned} \tag{6.5}$$

6.2.2 Computing the Distribution of the RF Distance with the Fast Fourier Transform

In this dynamic programming algorithm, the most expensive part is computing the two-dimensional convolution in part 4 of Lemma 4, which takes $O(n^4)$ time [26]. The two-dimensional convolution of two matrices x and y are defined as:

$$z(i_1, i_2) = \sum_{j_1} \sum_{j_2} x(j_1, j_2) y(n_1 - j_1, n_2 - j_2)$$

We improve the running time by using the Fast Fourier Transform (*FFT*) [138, 193] to compute the two-dimensional convolution, which takes $O(n^2\log(n))$. The *FFT* is an algorithm for computing the Discrete Fourier Transform (*DFT*) of a matrix, which efficiently converts it from the time domain to the frequency domain [138, 197]. The convolution theorem states that multiplication in the frequency domain corresponds to convolution in the time domain:

$$\mathcal{F}\{f * g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\},$$

where $\mathcal{F}$ denotes the Discrete Fourier transform and $*$ denotes the convolution operator [90]. We use the *DFT* to convert the input matrices from the time domain into the frequency domain. The discrete Fourier transform transforms a sequence of N numbers $\{x_0, x_1, \ldots, x_{N-1}\}$ into another sequence of complex numbers $\{X_0, X_1, \ldots, X_{N-1}\}$ which is defined by:

$$X_k = \sum_{j=0}^{N-1}(x_j e^{\frac{-2\pi i}{N}kj})$$

The DFT of a vector of length N can be computed by left-multiplying it by a Fourier matrix of a suitable dimension.

$$M = \begin{bmatrix} 1 & 1 & 1 & . & 1 \\ 1 & w & w^2 & . & w^{1\cdot(N-1)} \\ 1 & w^2 & w^4 & . & w^{2\cdot(N-1)} \\ . & . & . & . & . \\ 1 & w^{(N-1)\cdot 1} & w^{(N-1)\cdot 2} & . & w^{(N-1)\cdot(N-1)} \end{bmatrix}$$

In the *FFT*, the unit w is the N^{th} root of unity, $e^{-2\pi i/N}$. After transforming the data into the frequency domain, all the computations happen over the complex numbers. The procedure for computing the convolution of two vectors using *FFT* is:

- Compute the *DFT* of each vector using *FFT*;
- Perform the point-wise multiplication of the two preceding results;
- Apply the inverse *FFT* (using $w^{-1} := e^{2\pi i/N}$ instead of w) to this product.

Unfortunately, convolutions via the *FFT* cannot be performed numerically stably for large N due to rounding errors. This is overcome by using the Number-Theoretic Transform (*NTT*) instead.

6.2.3 Computing the Distribution of the RF Distance with the Number-Theoretic Transform

The Number-Theoretic Transform (*NTT*) [196] is a generalization of the *DFT*, with the only difference being that the w is replaced with an N^{th} root of unity modulo a prime p, for a transformation of length N. In other words, the transformation is done over the number field Z_p formed by the integers modulo the prime p instead of the complex numbers $\mathbb{C}$. The first step in *NTT* is to pick a suitable modulus $p = kN + 1$ for some integer k, so that all the elements of the input and output are less than the modulus (for the convolution of two vectors of length N with maximum value M, the prime p needs to be greater than the maximum output value $M^2 \cdot N$). Dirichlet's theorem [172] guarantees that, for any vector of length N defined in a finite field, there exists some k that results in a suitable modulus p. We then compute the transformation matrix by finding a suitable w such that $w^N = 1$. In order to find such a w, we find a generator r in Z_p, i.e. an element such that when x goes from 1 to $p - 1$, $r^x \mod p$ covers all the numbers $1, 2, \ldots, (p - 1)$ in some order. We then set $w := r^k \mod p$. Indeed, $w^N \equiv r^{kN} \equiv r^{p-1} \equiv 1 \mod p$ by Fermat's Little Theorem. The transformation matrix for computing the *NTT* is then computed by using this w in the matrix defined in the previous section.

Similarly to the *FFT*, replacing w with $w^{-1} \mod p$ results in the inverse transformation matrix which is used for an inverse *NTT*. Otherwise, the process is exactly the same as for the *FFT*. Since all the numbers in the *NTT* computation are integers, there is no rounding error. The *NTT* can be implemented using an *FFT* algorithm such as Cooley-Tukey [36]. In order to do a one dimensional *NTT* with $O(N \log N)$ complexity, the transformation length N has to be a power of two.

As the size of a tree increases, the input elements become large, and it is challenging to calculate a suitable modulus p for the convolution using *NTT*. A possible solution for this would be to calculate the convolution with multiple *NTT*s, each with a different modulus p, and get the final result using the Chinese Remainder Theorem [48]. Another solution is using the approximation proposed in [26] to compute the approximate distribution of the *RF* distance [26].

Since we are computing the distribution of the *RF* distance from a given tree, the numbers that are involved in our computations are always less than $b(n)$ (the number of unrooted labeled phylogenetic trees with n tips):

$$b(n) = (2n - 5)!!$$

From the above formula it is clear that these numbers grow exponentially with n. Using the asymptotic expression of Vaclav Kotesovec for odd numbers, we have this asymptotic approximation for $b(n)$ and accordingly for the involved numbers [2, 1]:

$$b(n) \sim \sqrt{2}(2n - 5)^{n-2} e^{-\frac{(2n-5)}{2}}$$

Having this approximation, the involved numbers can be written in $O(n \log(n))$ bits. So, if we count the bits, rather than the numbers, the complexity would be increased by a factor of $O(n \log(n))$, but it would not affect the whole time complexity since the time complexity of computing the convolution is $O(n^2 \log(n))$.

Theorem 1. *Given a binary tree T with n tips, we can compute the coefficients of the generating function $\mathcal{B}(T, x)$ in $O(n^3 \log n)$ time.*

Proof. If v has no children in $\mathcal{I}(T)$, then we can compute $R(v, s, k)$ in constant time based on Lemmas 3 and 4 (part 1). If v has one child in $\mathcal{I}(T)$, then the dominant term for computing $R(v, s, k)$ is equation 6.3, which can be computed in $O(n^2)$ time for all $s, k \leqslant n - 3$. Finally, if v has two children in $\mathcal{I}(T)$, then the most computationally expensive term for computing $R(v, s, k)$ is equation 6.5 which we compute it in $O(n^2 \log n)$ by leveraging *FFT* or the *NTT*. In order to compute all the coefficients $r_s(T)$, we need to compute $R(v, s, k)$ for all the internal nodes of T; Hence we can compute all the coefficients $r_s(T)$ in $O(n^3 \log n)$. Using equation (6.1) and the relation between $b_m(T)$ and $q_s(T)$, we can compute the $q_s(T)$'s, and finally the $b_m(T)$'s, in $O(n^3 \log n)$ time. Note that the time complexity of computing the convolution using *NTT* depends on the transformation length (N) and N increases as the number of tips increase. Since $N = O(n)$, the overall computation depends on n. □

6.3 Results

We implemented the dynamic programming approach introduced by Bryant and Steel [26], as well as our approach, in *R* [154] and Python. In order to compute the convolution, we first leveraged the *FFT* via the *convolve* command in the *stat* package in *R*. The *convolve* command performs a one-dimensional convolution using *FFT*; however, we have to compute a two-dimensional convolution on two input matrices. In order to account for this, we convert our two input matrices into vectors using sufficient zero-padding, apply a one-dimensional convolution, and remove the padding. We were able to generate the *RF* distance distribution for trees with 5 to 22 tips efficiently using *FFT*.

We were not able to compute the distribution on trees with more than 22 tips due to the numerical instability of the *FFT* applied to vectors containing numbers spanning many orders of magnitude. For example for this random tree with 26 tree tips:

$$(((((((,),((,),(,))),((,),)),(,)),(((,),),)),(((,((,),)),),(((,),),))),(,))$$

The range of the entries of $R(v, s, k)$ (v is one of the root's children) is between 1 to about 10^{16}. When we apply *convolve* command on two vectors with large range of entries, the results are some negative values which is impossible. We thus used the *NTT* to compute the convolution instead.

Our implementation in *R* works efficiently for small trees, up to $n = 50$ tips; we used the Cooley-Tukey algorithm [36] to implement the fast *NTT*. For larger trees, we used the *gmp* package [111] in *R* to perform the computation using large numbers. However, the code ran a lot slower when using the *gmp* package. Therefore, in order to compute the convolution with larger numbers, we made use of an efficient Python implementation of the *NTT* [128]. This enabled us to implement the dynamic programming algorithm [26] and our proposed algorithm in Python. Figure 6.4 shows the *RF* distance distribution for the completely balanced and unbalanced (caterpillar) trees with $8, 16, 32$ and 64 tips. Our Python and R packages are freely available at `https://github.com/WGS-TB/RFDistribution` and `https://github.com/WGS-TB/RFDistributionR` respectively.

We compute the distribution of the Robinson-Foulds distance using our method for two sets of extreme trees: completely balanced unrooted and completely unbalanced unrooted trees. Completely balanced unrooted trees can not be described easily because they are not unique. In this paper, we consider unrooted binary trees with most possible number of cherries located evenly through the tree as completely balanced unrooted trees (Figure 6.3(a)). The completely unbalanced unrooted tree or so-called unrooted caterpillar is the unique unrooted binary tree with just two cherries. (Figure 6.3(b))

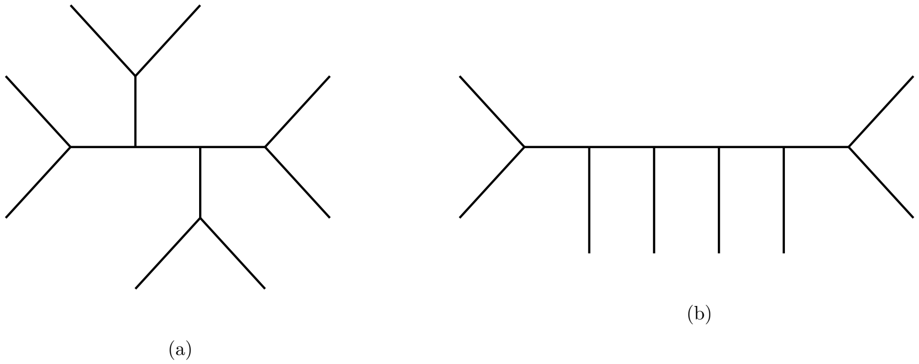


Figure 6.3: (a) shows the maximally balanced (completely balanced unrooted tree) with 8 tips, and (b) shows the unrooted caterpillar (completely unbalanced tree) on 8 tips

Computing a suitable prime for large trees (trees with $n \geq 200$) is computationally expensive, and we cannot use a simple *NTT* with our currently computed prime moduli to perform the convolution. A possible solution for this is to use the Chinese Remainder Theorem *(CRT)* [83].

We used our Python implementation to compare the algorithm with and without *NTT* (Figure 6.5) for a set of trees with 12 to 75 tips, in increments of 3. For each number of tips n, we only considered the trees with $n/3 - 1$ internal nodes of type III (internal nodes with two internal children), to ensure that the convolution is called the same number of times for all trees of the same size. For each n, we randomly generated 10 such trees and recorded the mean running time. Figure 6.5(a) shows the method's running time without *NTT*; the slope of the log-log plot of time vs. n is 5.035, close to the expected 5. Figure 6.5(b) shows the method's running time with *NTT*. Its slope is close to 3, but jumps when

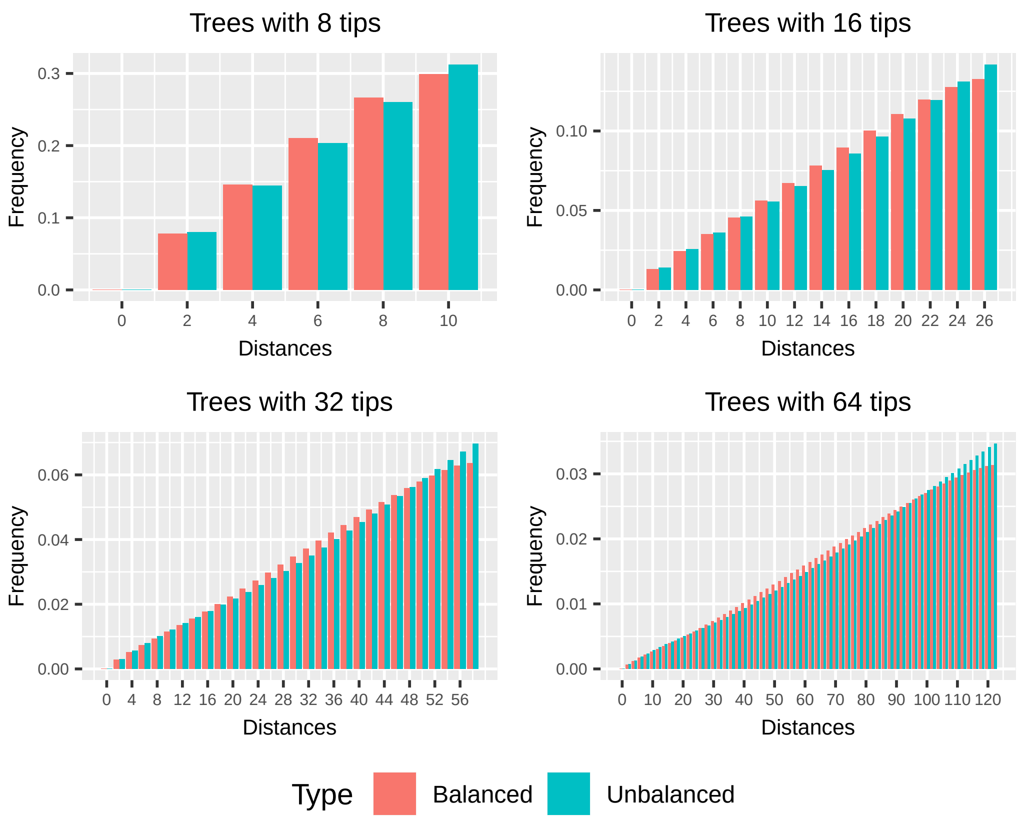


Figure 6.4: Log scale plot showing the RF distribution of the completely balanced and the completely unbalanced labeled trees with $8, 16, 32$ and 64 tips. The y-axis represents the frequency of the distances from a given phylogenetic tree.

the transformation length changes, as the running time of NTT-based convolutions depends on it linearly (Figure 6.6).

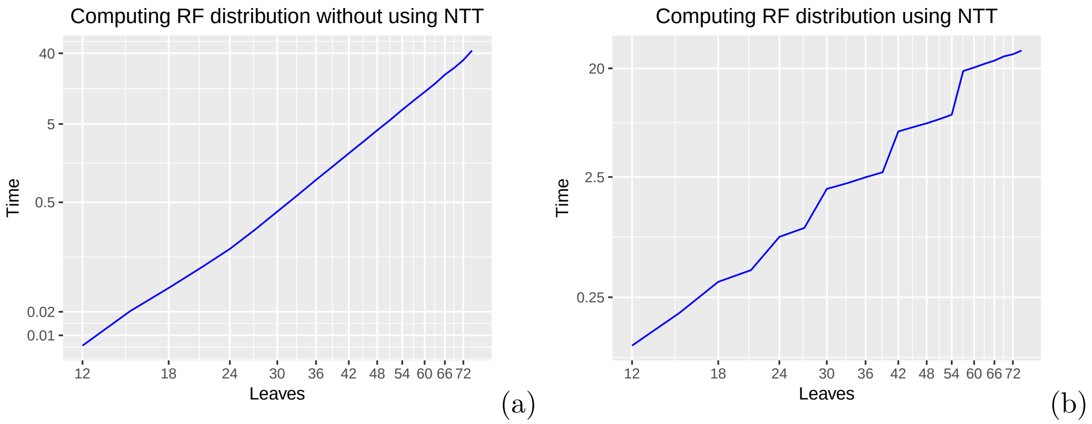


Figure 6.5: a) log-log plot showing the running time of the algorithm when we do not use the NTT to compute the convolution. The slope of this plot is 5.035, in agreement with the $O(n^5)$ running time of the algorithm. b) log-log plot showing the running time of the algorithm when we use NTT to compute the convolution. There are some jumps in the plot, indicating the points where the transformation length changed. This figure shows that using NTT to compute the convolution can improve the performance of the computation of the RF distance distribution.

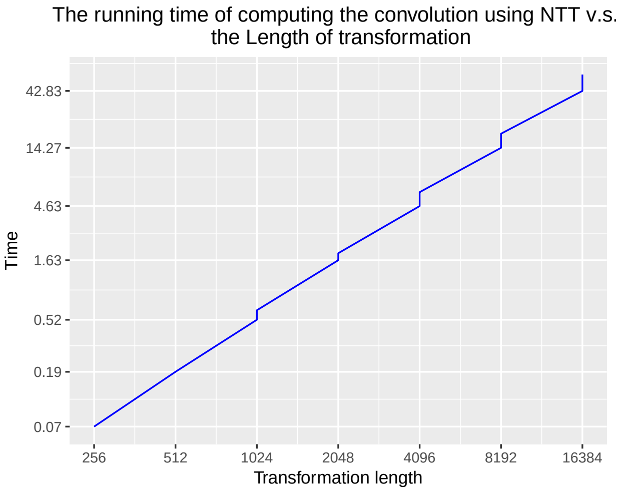


Figure 6.6: log-log plot showing the running time of computing the convolution using NTT vs. transformation length. The running time changes nearly linearly with the transformation length.

Chapter 7

Conclusions

Phylogenetic trees are most often used in biology to study the historical relationship between several species or organisms. These trees contain both branch lengths and information in the form of the tree topology or shape. A tree's shape specifies the connectivity of a tree, while its branch lengths reflect either the time or genetic distance between branching events. The Colless and Sackin indices are among the most well-known measures of tree shape. Tree shape statistics have a wide range of applications, from phylogenetic hypothesis testing to recent applications in phylodynamics. Phylogenetic trees have been used in infectious disease to estimate the basic reproduction number, parameters of transmission models, aspects of underlying contact networks, and in densely sampled datasets even person-to-person transmission events and timing. It is therefore natural to hypothesize that phylogenetic tree structures and branching patterns contain information about short-term growth and fitness. In this dissertation, we focus on introducing new tree shape statistics (tools for measuring the shape of a tree) and evaluating the previously proposed ones. We also investigate the problem of using machine learning tools and phylogenetic properties of the influenza tree to predict which influenza strains will persist into the future. Moreover, we improved the running time of the best known algorithm for computing the distribution of the Robinson-Fould distance for a given tree.

7.1 Summary

In chapter 1, we introduced phylogenetic trees and briefly explained some definitions and notations used in phylogenetics. We discussed the stochastic evolutionary models, including the Yule and the PDA models in this chapter. We also introduced tree shape statistics, which are tools to measure the degree of imbalance or asymmetry of a tree shape. Moreover, we had a brief look at methods for tree comparison.

In chapter 2, we reviewed and discussed the classical tree shape statistics and the recently proposed ones and also the methods proposed for evaluating the discriminatory power of the existing tree shape statistics. We also reviewed the applications of tree shape statistics in evolutionary hypothesis testing and the recent applications in phylodynamics. Finally, we discussed the methods for computing the distribution of the RF-distance for a given tree.

In chapter 3, we proposed a new resolution function based on the Laplacian matrix to evaluate different tree shape statistics. This resolution function can rank the statistics in terms of their power in discriminating all possible phylogenetic trees on the same number of leaves. Among our new resolution function and the previously proposed ones, the top statistics are $Colless$ index, $Sackin$ index, and our proposed statistic (their linear combination), and the worst ones are B_1 and I_2. The advantage of our new resolution function is to reduce the time and space complexity of the computation while producing comparable results. Being able to explore the set of trees with more tips, allows us to ensure that the trends observed with smaller trees are not artifactual, and remain when we explore larger trees.

In chapter 4, we proposed two classes of tree shape statistics. The first one is the linear combinations of two existing statistics that are optimal with respect to a resolution function, and show evidence that the statistics in this class converge to a limiting linear combination as the size of the tree increases. Using the geometric resolution function introduced in chapter 3, we showed that our proposed tree shape statistics Saless can perform better than classical tree shape statistics in distinguishing between dissimilar trees. The other class of proposed tree shape statistics is inspired by network science. We used some concepts from network science, namely, diameter, average path length, betweenness, closeness, and eigenvector centrality to define some new tree shape statistics to summarize the shape of a phylogenetic tree. Using mutual information and supervised learning algorithms, we showed that the statistics adapted from network science perform as well as or better than conventional statistics.

In chapter 5, we efficiently predict the success of individual influenza virus subtrees using machine learning tools applied to phylogenetic trees. Our method allows binary classifiers to be trained to predict which currently circulating subtrees will persist into the future based primarily on a suite of phylogenetic features. Our approach is complementary to previous approaches, including the fitness based model proposed by Łuksza and Lässig [112] and the tree-centred work of Neher and colleagues [131, 130]. Our approach requires a reconstructed timed phylogenetic tree, and can accommodate additional data (e.g. we have used epitope mutations) easily. Other approaches often require additional data such as HI titers and estimates of the ancestral sequences, introducing experimental and computational costs and uncertainty. Our approach makes use of the reconstructed phylogeny in two distinct ways. First, in obtaining the groups of taxa (subtrees) considered together for analysis, and second in that the tree shape and length features are derived from the phylogeny, and capture features of the complex branching patterns within subtrees, as opposed to their overall rates.

In chapter 6, we improved the running time of the algorithm introduced by Bryant and Steel [26] to compute the RF distance distribution of a given tree. Their dynamic programming algorithm has a time complexity of $O(n^5)$, and the dominant term is due to computing a two-dimensional matrix convolution, which takes $O(n^4)$ per iteration if done naïvely. Using the Number Theoretic Transform (NTT) for computing the convolution, we reduced the running time of the proposed algorithm to $O(n^3 \log n)$. We implemented their proposed algorithm and our proposed modification in both *R* statistical computing language [154] and Python. We compared the running time of our approach and the algorithm by Bryant and Steel [26] on a set of trees up to 75 tips and showed that the running time of the algorithm is in agreement with the theoretical analysis.

7.2 Discussion

In chapter 3, we have implemented our proposed method in the *R* statistical computing language [154]. The challenge of implementing the method was in handling large matrices, as the number of unlabeled trees n grows exponentially with the number of leaves n. Our implementation needs to allocate a vector of size n^2, which is not possible since *R* holds all objects in virtual memory, and each object can use a limited amount of memory. One of the advantages of using the Laplacian matrix in our method is its sparsity, which enables us to implement it via the *Matrix* package. We also use a specific numbering scheme for labeling the phylogenetic trees to account for tree isomorphism, which results in reduced time and memory requirements. A better implementation would allow us to extend distance and Laplacian matrix computations to larger tree sizes. An alternative approach is to reach larger tree sizes by replacing the exact computation that we pursue here with a Monte Carlo Markov Chain approach, which is feasible because the neighbors of each tree with respect to a rearrangement distance can be readily produced.

In chapter 5, we proposed a method to predict the successful influenza strains that would persist in the future outbreak. Influenza virus A can be categorized based on the presence of different proteins on the surface of the viruses: hemagglutinin (HA) and neuraminidase (NA). We use trees reconstructed from HA sequences; relatedness in the HA tree corresponds to similar HA sequences and hence to similar immunity profiles, as antibodies are induced by HA. Indeed, path lengths in the HA phylogenetic tree provide a good model for antigenic differences modeled by serological assays [131]. Trees describe the relative number of recent descendants of a lineage compared to closely-related lineages, the timing and asymmetry in the descent patterns and the short- and long-term future populations that are related to the lineages. Our approach allows this information to be included in predictive efforts.

We used RAxML to infer the trees; it uses a maximum likelihood approach and is considered a state-of-the art reconstruction algorithm [205, 100]. Due to the large numbers of isolates, we did not perform Bayesian Markov Chain Monte Carlo (MCMC) tree reconstruction to accommodate tree uncertainty; in addition to each required MCMC run, this would also have required exploration of different priors and assumptions, and it is computationally unfeasible for thousands of tips. In order to check the robustness of our approach, we used different training and testing trees, including training on H3N2 but testing on H1N1, pooling H3N2 and H1N1 and using distinct time slices and consistently obtained successful predictions.

Tree shape statistics are dependent on the size of a subtree, so in order to have a more robust comparison between the subtrees, it would be best to select subtrees with approximately the same size. In our data, the size of a trimmed subtree was a poor predictor of the fractional growth. Furthermore, constraining the sizes of subtrees reduces both the number of subtrees and the number of tips that can be included in the analysis, and size may in fact be a valuable predictor. We chose the approach here to balance these contrasting issues; the ongoing sampling and sequencing and the natural passing of time will ultimately provide more data – more years, and more samples per year – such that subtrees of more consistent size can be used. We anticipate that this will increase the quality of predictions. Another next logical step will be to model competition or other interactions between major lineages. We have not explicitly modeled ancestral states, key individual mutations, serological data or estimates of the fitness of sequences, but our approach could easily integrate results from models that include these features. Ideally, all relevant sources of information would be integrated and updated in real time [130, 112]. However, while short-term forecasting based on various data sources is feasible and is required to update seasonal vaccines, perfect short-term prediction and accurate long-term prediction are likely not possible because evolutionary events are fundamentally stochastic.

The potential overlap between subtrees could induce dependence in the outcome variable (success), i.e., if nodes n_i and n_j have overlap, and n_i is successful, then n_j may be more likely to be successful. Notice that having some tips in common does not mean that overall subtree features are similar, but we hypothesize that the chance increases as the proportion of shared tips grows. If n_i was in the training data and n_j in the test data, and if the two subtrees had similar features, then correlations in the success of these two subtrees could result in overfitting the data. Controlling this potential effect is one reason to use a cutoff of 3.4 years for success (meaning that n_j could be unsuccessful with n_i successful). In one of our experiments, we trained on H3N2 and tested on H1N1 (see Figure 5.2(b)) to ensure that test and training data are completely distinct. We also explored the performance of our approach on a set of pool subtrees from both H1N1 and H3N2.

Our data are censored, because we cannot observe the future of subtrees beginning in 2018; furthermore, we have limited knowledge of the true success of subtrees beginning in the most recent 3.4 years of our data (since it takes approximately 2 years following the end of the trimmed subtree before its success is known). We only know whether a subtree has been successful if it has already had a sufficient number of tips; other subtrees may yet do so. Accordingly, we could not train our models using the last few years of data. We used 10-fold cross-validation on the set of subtrees whose ancestral nodes arose before 2015-1. We refer to this as the set of "non-recent subtrees". This results in one "out of fold" prediction for each such subtree. We trained an additional "general" model on the set of non-recent subtrees. We then had 10 cross-validation models and one additional model that we could use to make predictions on the subtrees arising after 2015 (for which the true success is only partially known).

The structure (and hence internal nodes) of a large phylogeny depends on all tips, not only the tips prior to the node chronologically. To avoid having the "future" tips affect the nodes on which we based our analysis, we also used a "time slicing" approach that is amenable to use in real-time, season by season. Here, we extracted subtrees not from the full phylogeny but from a tree containing only tips prior to a fixed time. We then assessed the success of these subtrees with reference to later tips (see Figures 5.3 and 5.7).

Direct comparison of our method with previous approaches is challenging as different approaches have various definitions of success and use different evaluation metrics. Neher et al. [131, 130] demonstrate that the shape of the reconstructed phylogenetic trees contains information about the fitness of the sample sequences and using this information, they proposed a model to predict the successful strains in the upcoming influenza season, using the local branching index. The units of prediction in their approach are strains while we use subtrees. One of the advantages of using subtrees over strains is that strains are typically observed only in a single season, while subtrees have an evolutionary history. Our tests have found that at the subtree level, local branching index does not perform as well as our combined feature sets. Łuksza et al. [112] infer the fitness components of influenza strains: adaptive epitope changes and deleterious mutations outside the epitopes for the strains circulating in a given year, using population-genetic data of all previous strains. They then used the fitness and frequency of each strain to predict the frequency of its descendent strains in the following year. They used clades as units of prediction, which are close to our units of prediction (subtrees). However, we put some time restriction on choosing the subtrees, which enable us to compare subtrees in the same time interval. In order to have a more balanced dataset and accordingly, a more powerful model, we choose the subtree growth ratio equal to 1.1 and predict the growth of the subtrees in 2 years. These parameters are 1 and 1 year, respectively, in the model proposed by Łuksza et al. We do not have strain frequency data, but a comparison with epitope features as the primary basis of prediction found that our feature sets performed better.

We did not include information about proximity of strains to current or recent vaccines, which might have led to false positives in our results, if a subtree showed early signs of success but was later suppressed by vaccination (but this is unlikey, as only a small portion of the global population is vaccinated). We also did not explicitly include immunological assay data, as these are not generally available. We do not have good estimates of the current frequencies of subtrees or strains - indeed, if up-to-date global frequencies were available at high resolution it would greatly facilitate short-term prediction. We used epitope sites following the approach of [112]; a model reflecting the impact of polymorphisms across more locations in HA and in other genes, if this were available, might improve predictions.

Recent studies on the viral isolates from vaccinated individuals indicate that they are significantly distinct from the vaccine strain and are broadly distributed on the tree, resulting in accelerated antigenic evolution [186]. Researchers have been working to develop a universal vaccine that would provide broad protection against both seasonal and pandemic influenza. Recent studies have also indicate that universal vaccines could decelerate the speed of evolution [126, 186]. If successful, such universal vaccines would eliminate the need for continual updates to seasonal influenza virus vaccines, but we would suggest that even so, current efforts to make short-term predictions based on surveillance and sequence data gathered over time can yield both practical results and broader insights into short-term patterns of evolution. Our approach indicates that fitness can affect the phylogenetic properties of a tree reconstructed from influenza viruses and that properties of small subtrees can be used as a set of predictors to estimate which groups of sequences are showing signs of success.

In chapter 6, we proposed using Number Theoric Transform (NTT) to compute a two-dimensional convulsion. To our knowledge, our method represents one of the rare times that the NTT is used to solve a computational biology problem, the last example we were able to find [14] dating to 1990. We believe that many similar applications, where numerical stability issues make the use of a FFT impractical, can be found. We chose the Robinson-Foulds (RF) distance since it is one of the most widely used measure of dissimilarity between trees. One of its main advantages is its time complexity, which can be computed in linear time. However, the RF distance is overly sensitive to some small changes in the tree. For example, just moving a leaf at the end of a caterpillar tree to the other end will result in a tree with the maximum possible RF distance to the original tree. Our results in accordance with the previous results show that the RF distance between two random binary trees with n tips has a very skewed distribution in which most values equal $n - 3$ [108].

7.3 Future Work

In chapter 3, a better implementation would allow us to extend the distance and Laplacian matrix computations to larger tree sizes. An alternative approach is to reach larger tree sizes by replacing the exact computation that we pursue here with a Monte Carlo Markov Chain approach, which is feasible because the neighbors of each tree with respect to a rearrangement distance can be readily produced.

In chapter 4, we investigated the optimal combination of different pairs of tree shape statistics. We conjecture that λ values converge for any pair of reasonable statistics. We cannot make any conclusion based on the small trees we examined so far, since convergence is a long-term behavior, and we leave the proof of this conjecture for future work. Regarding the application of tree shape statistics to phylodynamics, more powerful statistics, such as the pairwise combinations we introduced, are clearly needed. Tree shape statistics are used, for instance, as the features in predictive models of short-term influenza evolution and fitness models [131, 74]. Using more highly resolving features would arguably result in more accurate predictions. An additional future research direction could then be the extension of optimal combinations to more than 2 statistics, in which case one would need to optimize multiple coefficients at once.

In chapter 5, our approach is rooted in the hypothesis that fitness and early success leave signatures in the branching time and structure of phylogenetic trees, which can be complemented with additional relevant information such as epitope diversification. With only slight modifications, our model could be applied to other organisms. We could also extend the approach and train regression models to predict the number of tips arising from subtrees. However, a natural limitation of this (and other tree-based approaches) is that it detects signs of early growth - if an adaptive new mutation arose in the population and was sampled before that early growth could occur and be sampled, then we could not detect early signs of growth and would not see the new adaptive mutation. In contrast, a principled modeling approach based on an understanding of both what makes an influenza virus fit and on the current composition of population immunity might be able to detect fit novel mutations without relying on such viruses already have begun to spread.

In chapter 6, in order to compute the RF-distance distribution of a large tree (with $n \sim 10^2 - 10^4$), one would need to use the Chinese Remainder Theorem *(CRT)*, as computing a suitable prime modulus for computing the RF distance distribution of a larger tree is computationally expensive. We propose this as a future direction.

Bibliography

[1] Bounds of double factorial. https://math.stackexchange.com/questions/2832468/bounds-of-double-factorial.

[2] The online encyclopedia of integer sequences. http://oeis.org/A006882. Accessed: 2014-11-8.

[3] Milton Abramowitz and Irene A Stegun. Handbook of mathematical functions with formulas, graphs, and mathematical tables, volume 55. US Government printing office, 1948.

[4] Paul-Michael Agapow and Andy Purvis. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. Systematic biology, 51(6):866–872, 2002.

[5] David Aldous. Probability distributions on cladograms. In Random discrete structures, pages 1–18. Springer, 1996.

[6] David Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Statistical science, pages 23–34, 2001.

[7] Benjamin L Allen and Mike Steel. Subtree transfer operations and their induced metrics on evolutionary trees. Annals of combinatorics, 5(1):1–15, 2001.

[8] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. Science, 286(5439):509–512, 1999.

[9] Joëlle Barido-Sottani, Veronika Bošková, Louis Du Plessis, Denise Kühnert, Carsten Magnus, Venelin Mitov, Nicola F Müller, Jūlija PečErska, David A Rasmussen, Chi Zhang, et al. Taming the beast—a community teaching material resource for beast 2. Systematic biology, 67(1):170–174, 2017.

[10] Ian G Barr, Colin Russell, Terry G Besselaar, Nancy J Cox, Rod S Daniels, Ruben Donis, Othmar G Engelhardt, Gary Grohmann, Shigeyuki Itamura, Anne Kelso, et al. WHO recommendations for the viruses used in the 2013–2014 northern hemisphere influenza vaccine: Epidemiology, antigenic and genetic characteristics of influenza A(H1N1)pdm09, A(H3N2) and B influenza viruses collected from october 2012 to january 2013. Vaccine, 32(37):4713–4725, 2014.

[11] Trevor Bedford and Richard Neher. Seasonal influenza circulation patterns and projections for Feb 2018 to Feb 2019. bioRxiv, 2018.

[12] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. Nucleic Acids Research, 33:D34–D38, 2005.

[13] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. Nucleic acids research, 41(D1):D36–D42, 2012.

[14] Donald C Benson. Digital signal processing methods for biosequence comparison. Nucleic acids research, 18(10):3001–3001, 1990.

[15] Elisabetta Bergamini, Michele Borassi, Pierluigi Crescenzi, Andrea Marino, and Henning Meyerhenke. Computing top-k closeness centrality faster in unweighted graphs. ACM Transactions on knowledge discovery from data (TKDD), 13(5):53, 2019.

[16] Michael GB Blum and Olivier François. On statistical tests of phylogenetic tree imbalance: the sackin and other indices revisited. Mathematical biosciences, 195(2):141–153, 2005.

[17] Michael GB Blum and Olivier François. Which random processes describe the tree of life? a large-scale study of phylogenetic tree imbalance. Systematic biology, 55(4):685–691, 2006.

[18] Michael GB Blum, Olivier François, and Svante Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. Annals of applied probability, pages 2195–2214, 2006.

[19] Béla Bollobás. Modern Graph Theory. Springer, 2013.

[20] Nicolas Bortolussi, Eric Durand, Michael Blum, and Olivier Francois. apTreeshape: Analyses of phylogenetic treeshape, 2012. R package version 1.4-5.

[21] Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. PLoS computational biology, 15(4):e1006650, 2019.

[22] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In ACM Sigmod record, volume 29, pages 93–104. ACM, 2000.

[23] Gerth Stølting Brodal, Rolf Fagerberg, and Christian NS Pedersen. Computing the quartet distance between evolutionary trees in time o (n log n). Algorithmica, 38(2):377–395, 2004.

[24] Christopher Brown. hash: Full feature implementation of hash/associated arrays/dictionaries, 2013. R package version 2.2.6.

[25] Daniel G Brown and Jakub Truszkowski. Fast phylogenetic tree reconstruction using locality-sensitive hashing. In International Workshop on Algorithms in Bioinformatics, pages 14–29. Springer, 2012.

[26] David Bryant and Mike Steel. Computing the distribution of a tree metric. IEEE/ACM Transactions on computational biology and bioinformatics, 6(3):420–426, July 2009.

[27] David Bryant, John Tsang, Paul Kearney, and Ming Li. Computing the quartet distance between evolutionary trees. Symposium on discrete algorithms, 11 1999.

[28] Peter Buneman. The recovery of trees from measures of dissimilarity. Mathematics in the archaeological and historical sciences, 1971.

[29] Scott Chasalow. combinat: combinatorics utilities, 2012. R package version 0.0-8.

[30] Leonid Chindelevitch, Maryam Hayati, Art F. Y. Poon, and Caroline Colijn. Network science inspires novel tree shape statistics. bioRxiv doi:10.1101/608646, 2019.

[31] Caroline Colijn and Jennifer Gardy. Phylogenetic tree shapes resolve disease transmission patterns. Evolution, medicine, and public health, 2014(1):96–108, 2014.

[32] Caroline Colijn and Jennifer Gardy. Phylogenetic tree shapes resolve disease transmission patterns. Evol Med Public Health, 2014(1):96–108, 9 June 2014.

[33] Caroline Colijn and Giacomo Plazzotta. A metric on phylogenetic tree shapes. Systematic biology, 67:113–126, 2018.

[34] Donald H Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. Systematic zoology, 31(1):100–104, 1982.

[35] Donald H Colless. Relative symmetry of cladograms and phenograms: an experimental study. Systematic biology, 44(1):102–108, 1995.

[36] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex Fourier series. Mathematics of computation, 19(90):297–301, 1965.

[37] Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[38] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. InterJournal, Complex Systems:1695, 2006.

[39] Katalin Csilléry, Michael G.B. Blum, Oscar E. Gaggiotti, and Olivier François. Approximate Bayesian Computation (ABC) in practice. Trends in ecology & evolution, 25(7):410 – 418, 2010.

[40] Karel Culik II and Derick Wood. A note on some tree similarity measures. Information processing letters, 15(1):39–42, 1982.

[41] Bhaskar DasGupta, Xin He, Tao Jiang, Ming Li, John Tromp, and Louxin Zhang. On computing the nearest neighbor interchange distance. In DIMACS workshop on discrete problems with medical applications, volume 55, page 125. American mathematical Soc., Press, 2000.

[42] William HE Day. Properties of the nearest neighbor interchange metric for trees of small size. Journal of theoretical biology, 101(2):275–288, 1983.

[43] Adel Dayarian and Boris I Shraiman. How to infer relative fitness from a sample of genomic sequences. Genetics, 197(3):913–923, 2014.

[44] Alexandre De Bruyn, Darren P Martin, and Pierre Lefeuvre. Phylogenetic reconstruction methods: an overview. In Molecular plant taxonomy, pages 257–277. Springer, 2014.

[45] Damien M de Vienne, Tatiana Giraud, and Olivier C Martin. A congruence index for testing topological similarity between trees. Bioinformatics, 23(23):3119–3124, 2007.

[46] Xavier Didelot, Christophe Fraser, Jennifer Gardy, and Caroline Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Molecular biology and evolution, 34(4):997–1007, 2017.

[47] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, Andreas Weingessel, and Maintainer Friedrich Leisch. The e1071 package. Misc functions of department of statistics (e1071), TU Wien, 2006.

[48] Pei Dingyi, Salomaa Arto, and Ding Cunsheng. Chinese remainder theorem: applications in computing, coding, cryptography. World Scientific, 1996.

[49] Dan E Dudgeon and Russell M Mersereau. Multidimensional Digital Signal Processing. Prentice-Hall Signal Processing Series. Prentice-Hall, Englewood Cliffs, NJ, 1984.

[50] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1998.

[51] Roger C Entringer, Douglas E Jackson, and DA Snyder. Distance in graphs. Czechoslovak mathematical journal, 26(2):283–296, 1976.

[52] Paul Erdos. On random graphs. Publicationes mathematicae, 6:290–297, 1959.

[53] Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution, 17(6):368–376, 1981.

[54] Joseph Felsenstein and Joseph Felenstein. Inferring phylogenies, volume 2. Sinauer Associates Sunderland, 2004.

[55] Joseph Felsenstein, Stanley Sawyer, and Rochelle Kochin. An efficient method for matching nucleic acid sequences. Nucleic acids research, 10(1):133–139, 1982.

[56] Miroslav Fiedler. Algebraic connectivity of graphs. Czechoslovak mathematical journal, 23(2):298–305, 1973.

[57] Walter M Fitch and Emanuel Margoliash. Construction of phylogenetic trees. Science, 155(3760):279–284, 1967.

[58] Tomas Flouri, F Izquierdo-Carrasco, Diego Darriba, Andre J Aberer, L-T Nguyen, BQ Minh, Arndt Von Haeseler, and Alexandros Stamatakis. The phylogenetic likelihood library. Systematic biology, 64(2):356, 2015.

[59] Brian Thomas Foley, Bette Tina Marie Korber, Thomas Kenneth Leitner, Cristian Apetrei, Beatrice Hahn, Ilene Mizrachi, James Mullins, Andrew Rambaut, and Steven Wolinsky. HIV Sequence Compendium 2013. Technical Report LA-UR 13-26007, Los Alamos National Laboratory, NM, 2013.

[60] Simon DW Frost and Erik M Volz. Modelling tree shape and structure in viral phylodynamics. Philosophical transactions of the royal society B: Biological sciences, 368(1614):20120208, 2013.

[61] George W Furnas. The generation of random, binary unordered trees. Journal of classification, 1(1):187–233, 1984.

[62] Giuseppe Fusco and Quentin CB Cronk. A new method for evaluating the shape of large phylogenies. Journal of theoretical biology, 175(2):235–243, 1995.

[63] Douglas J Futuyma and Shawn S McCafferty. Phylogeny and the evolution of host plant associations in the leaf beetle genus ophraella (coleoptera, chrysomelidae). Evolution, 44(8):1885–1913, 1990.

[64] Olivier Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Molecular biology and evolution, 14(7):685–695, 1997.

[65] Chris Godsil and Gordon Royle. Algebraic Graph Theory. Springer, 2001.

[66] Pablo A. Goloboff, Joan S. Arias, and Claudia A. Szumik. Comparing tree shapes: beyond symmetry. Zoologica scripta, 46(5):637–648, 2017.

[67] Gene H Golub and Charles F Van Loan. Matrix computations. JHU Press, 3 edition, 2012.

[68] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James L N Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. Science, 303(5656):327–32, Jan 2004.

[69] Stephen Guattery and Gary L Miller. Graph embeddings and Laplacian eigenvalues. SIAM journal on matrix analysis and applications, 21(3):703–723, 2000.

[70] Stéphane Guindon, Frédéric Delsuc, Jean-François Dufayard, and Olivier Gascuel. Estimating maximum likelihood phylogenies with phyml. In Bioinformatics for DNA sequence analysis, pages 113–137. Springer, 2009.

[71] Craig Guyer and Joseph B. Slowinski. Adaptive radiation and the topology of large phylogenies. Evolution, 47(1):253–263, 1993.

[72] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 2018.

[73] EF Harding. The probabilities of rooted tree-shapes generated by random bifurcation. Advances in applied probability, 3(1):44–77, 1971.

[74] Maryam Hayati, Priscila Biller, and Caroline Colijn. Predicting the short-term success of human influenza a variants with machine learning. bioRxiv doi:10.1101/609248, 2019.

[75] Maryam Hayati, Bita Shadgar, and Leonid Chindelevitch. A new resolution function to evaluate tree shape statistics. PLoS One, 14(11): e0224197, 2019.

[76] Michael Hendy, CHC Little, and David Penny. Comparing trees with pendant vertices labelled. SIAM journal on applied mathematics, 44(5):1054–1065, 1984.

[77] Glenn Hickey, Frank Dehne, Andrew Rau-Chaplin, and Christian Blouin. Spr distance computation for unrooted trees. Evolutionary bioinformatics, 4:EBO–S419, 2008.

[78] Mark Holder and Paul O Lewis. Phylogeny estimation: traditional and bayesian approaches. Nature reviews genetics, 4(4):275, 2003.

[79] Katharina Huber, Andreas Spillner, Rados Suchecki, and Vincent Moulton. Metrics on multilabeled trees: Interrelationships and diameter bounds. IEEE/ACM Transactions on computational biology and bioinformatics (TCBB), 8(4):1029–1040, Jul. 2011.

[80] John P Huelsenbeck, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. Science, 294(5550):2310–2314, 2001.

[81] Gillian Hunt, Johanna Ledwaba, Anna Salimo, Monalisa Kalimashe, Beverly Singh, Adrian Puren, and Lynn Morris. Surveillance of transmitted HIV-1 drug resistance in 5 provinces in South Africa in 2011. Communicable diseases surveillance bulletin, 11:122–124, 2013.

[82] Hsien-Kuei Hwang and Ralph Neininger. Phase change of limit laws in the quicksort recurrence under varying toll functions. SIAM journal on computing, 31(6):1687–1722, 2002.

[83] Kenneth Ireland and Michael Rosen. A classical introduction to modern number theory, volume 84 of Graduate Texts in Mathematics. Springer science & business media, 2013.

[84] JP Jarvis, John K Luedeman, and Douglas R Shier. Comments on computing the similarity of binary trees. Journal of theoretical biology, 100(3):427–433, 1983.

[85] JP Jarvis, John K Luedeman, and Douglas R Shier. Counterexamples in measuring the distance between binary trees. Mathematical social sciences, 4(3):271–274, 1983.

[86] Lavanya Kannan and Ward C Wheeler. Maximum parsimony on phylogenetic networks. Algorithms for Molecular Biology, 7(1):9, 2012.

[87] Kazutaka Katoh, George Asimenos, and Hiroyuki Toh. Multiple alignment of dna sequences with mafft. Bioinformatics for DNA sequence analysis, pages 39–64, 2009.

[88] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic acids research, 30(14):3059–3066, 2002.

[89] Leo Katz. A new status index derived from sociometric analysis. Psychometrika, 18(1):39–43, 1953.

[90] Yitzhak Katznelson. An introduction to harmonic analysis. Cambridge University Press, 2004.

[91] Michelle Kendall, Michael Boyd, and Caroline Colijn. phyloTop: calculating topological properties of phylogenies, 2018. R package version 2.1.1.

[92] John Frank Charles Kingman. The coalescent. Stochastic processes and their applications, 13(3):235–248, 1982.

[93] Mark Kirkpatrick and Montgomery Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution, pages 1171–1181, 1993.

[94] Don Klinkenberg, Jantien A. Backer, Xavier Didelot, Caroline Colijn, and Jacco Wallinga. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLOS computational biology, 13(5), 2017.

[95] Teuvo Kohonen. Learning vector quantization. In Self-organizing maps, pages 175–189. Springer, 1995.

[96] Sergei L. Kosakovsky Pond, David Posada, Eric Stawiski, Colombe Chappey, Art F.Y. Poon, Gareth Hughes, Esther Fearnhill, Mike B. Gravenor, Andrew J. Leigh Brown, and Simon D.W. Frost. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLOS computational biology, 5(11):1–21, 11 2009.

[97] Mirko Křivánek. Computing the nearest neighbor interchange metric for unlabeled binary trees is NP-complete. Journal of classification, 3(1):55–60, 1986.

[98] Joseph B Kruskal and Myron Wish. Multidimensional scaling, volume 11. Sage, 1978.

[99] Max Kuhn, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. caret: Classification and Regression Training, 2017. R package version 6.0-78.

[100] John A Lees, Michelle Kendall, Julian Parkhill, Caroline Colijn, Stephen D Bentley, and Simon R Harris. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome open research, 3, 2018.

[101] Leonid Chindelevitch, Art F. Y. Poon, and Caroline Colijn. treeCentrality: A package for computing the network science statistics on trees in linear time. `https://rdrr.io/github/Leonardini/treeCentrality/man/treeCentrality.html`, July 2018. Accessed: 2019-3-28.

[102] Gabriel E Leventhal, Alison L Hill, Martin A Nowak, and Sebastian Bonhoeffer. Evolution and emergence of infectious diseases in theoretical and real-world networks. Nature communications, 6:6101, 2015.

[103] Gabriel E Leventhal, Roger Kouyos, Tanja Stadler, Viktor Von Wyl, Sabine Yerly, Jürg Böni, Cristina Cellerai, Thomas Klimkait, Huldrych F Günthard, and Sebastian Bonhoeffer. Inferring epidemic contact structure from phylogenetic trees. PLoS computational biology, 8(3):e1002413, 2012.

[104] Ted G. Lewis. Network Science: Theory and Applications. Wiley, 2009.

[105] Eric Lewitus and Helene Morlon. Characterizing and comparing phylogenies from their Laplacian spectrum. Systematic biology, 65(3):495–507, 2016.

[106] Gang Li. Generation of rooted trees and free trees. Master's thesis, University of Victoria, 1996. `http://webhome.cs.uvic.ca/~ruskey/Theses/GangLiMScThesis.pdf`.

[107] Ming Li, John Tromp, and Louxin Zhang. On the nearest neighbour interchange distance between evolutionary trees. Journal of theoretical biology, 182(4):463–467, 1996.

[108] Yu Lin, Vaibhav Rajan, and Bernard ME Moret. A metric for phylogenetic trees based on matching. IEEE/ACM transactions on computational biology and bioinformatics, 9(4):1014–1022, 2011.

[109] Yu Lin, Vaibhav Rajan, and Bernard ME Moret. A metric for phylogenetic trees based on matching. IEEE/ACM Transactions on computational biology and bioinformatics (TCBB), 9(4):1014–1022, 2012.

[110] Andrzej Lingas, Hans Olsson, and Anna Östlin. Efficient merging, construction, and maintenance of evolutionary trees. In International colloquium on automata, languages, and programming, pages 544–553. Springer, 1999.

[111] Antoine Lucas, Immanuel Scholz, Rainer Boehme, Sylvain Jasson, and Martin Maechler. gmp: Multiple Precision Arithmetic, 2018. R package version 0.5-13.2.

[112] Marta Łuksza and Michael Lässig. A predictive fitness model for influenza. Nature, 507(7490):57–61, 2014.

[113] Marc Manceau, Amaury Lambert, and Hélène Morlon. Phylogenies support out-of-equilibrium models of biodiversity. Ecology letters, 18(4):347–356, 2015.

[114] Kanti V Mardia. Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3):519–530, 1970.

[115] Frederick Matsen. A geometric approach to tree shape statistics. Systematic biology, 55(4):652–661, 2006.

[116] Frederick Matsen. Optimization over a class of tree shape statistics. IEEE/ACM Transactions on computational biology and bioinformatics (TCBB), 4(3):506–512, 2007.

[117] Rosemary M McCloskey, Richard H Liang, and Art FY Poon. Reconstructing contact network parameters from viral phylogenies. Virus evolution, 2(2), 2016.

[118] Andy McKenzie and Mike Steel. Distributions of cherries for two models of trees. Mathematical biosciences, 164(1):81–92, 2000.

[119] Cornelia Metzig and Caroline Colijn. Preferential attachment in systems and networks of constant size. arXiv:1811.04972, 2018.

[120] Charles D Michener and Robert R Sokal. A quantitative approach to a problem in classification. Evolution, 11(2):130–162, 1957.

[121] Arnau Mir, Francesc Rosselló, and Lucia Rotger. A new balance index for phylogenetic trees. Mathematical biosciences, 241(1):125–136, 2013.

[122] Charles Mitter, Brian Farrell, and Douglas J Futuyma. Phylogenetic studies of insect-plant interactions: insights into the genesis of diversity. Trends in ecology and evolution, 6(9):290–293, 1991.

[123] Bojan Mohar and Tomaž Pisanski. How to compute the wiener index of a graph. Journal of mathematical chemistry, 2:267–277, 1988.

[124] Arne O Mooers and Stephen B Heard. Inferring evolutionary process from phylogenetic tree shape. The quarterly review of biology, 72(1):31–54, 1997.

[125] Hélène Morlon, Eric Lewitus, Fabien L Condamine, Marc Manceau, Julien Clavel, and Jonathan Drury. Rpanda: an R package for macroevolutionary analyses on phylogenetic trees. Methods in ecology and evolution, 7(5):589–597, 2016.

[126] Dylan H Morris, Katelyn M Gostic, Simone Pompei, Trevor Bedford, Marta Łuksza, Richard A Neher, Bryan T Grenfell, Michael Lässig, and John W McCauley. Predictive modeling of influenza shows the promise of applied evolutionary biology. Trends in microbiology, 26(2):102–118, 2018.

[127] Niranjan Nagarajan, Neil Jones, and Uri Keich. Computing the *p*-value of the information content from an alignment of multiple sequences. Bioinformatics, 21(suppl_1):i311–i318, 2005.

[128] Project Nayuki. Number-Theoretic Transform (integer DFT), 2017. url: https://www.nayuki.io/page/number-theoretic-transform-integer-dft, [Online; update: 2017-06-07].

[129] Richard A Neher and Trevor Bedford. nextflu: Real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics, 31(21):3546–3548, 2015.

[130] Richard A Neher, Trevor Bedford, Rodney S Daniels, Colin A Russell, and Boris I Shraiman. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proceedings of the national academy of sciences, 113(12):E1701–E1709, 2016.

[131] Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. Elife, 3:e03568, 2014.

[132] Mark Newman. Analysis of weighted networks. Physical review E, 70(5):056131, 2004.

[133] Mark Newman. Networks. Oxford university press, 2018.

[134] Mark Newman, Albert-László Barabási, and Duncan J. Watts. The structure and dynamics of networks. The Princeton press, 2006.

[135] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution, 32(1):268–274, 2015.

[136] Melissa M Norström, Mattia CF Prosperi, Rebecca R Gray, Annika C Karlsson, and Marco Salemi. Phylotempo: a set of r scripts for assessing and visualizing temporal clustering in genealogies inferred from serially sampled viral sequences. Evolutionary bioinformatics online, 8:261, 2012.

[137] Vladimir Novitsky, Hermann Bussmann, Andrew Logan, Sikhulile Moyo, Erik van Widenfelt, Lillian Okui, Mompati Mmalane, Jeannie Baca, Lauren Buck, Eleanor Phillips, et al. Phylogenetic relatedness of circulating HIV-1C variants in mochudi, botswana. Plos One, 8(12):e80589, 2013.

[138] Henri J Nussbaumer. Fast Fourier transform and convolution algorithms, volume 2. Springer Science & Business Media, 2012.

[139] Roderic DM Page. On describing the shape of rooted and unrooted trees. Cladistics, 9(1):93–99, 1993.

[140] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: analyses of phylogenetics and evolution in R language. Bioinformatics, 20(2):289–290, 2004.

[141] Emanuel Parzen. Modern probability theory and its applications. John Wiley & Sons, Incorporated, 1960.

[142] Nicholas D Pattengale, Eric J Gottlieb, and Bernard ME Moret. Efficiently computing the robinson-foulds metric. Journal of computational biology, 14(6):724–735, 2007.

[143] Chuang Peng. Distance based methods in phylogenetic tree construction. Neural parallel and scientific computations, 15(4):547, 2007.

[144] Iosif Pinelis. Evolutionary models of phylogenetic trees. Proceedings of the royal society of London B: Biological sciences, 270(1522):1425–1431, 2003.

[145] Simone Pompei, Vittorio Loreto, and Francesca Tria. Phylogenetic properties of rna viruses. Plos One, 7(9):e44849, 2012.

[146] Art FY Poon. Phylodynamic inference with kernel ABC and its application to HIV epidemiology. Molecular biology and evolution, 32(9):2483–2495, 2015.

[147] Art FY Poon, Lorne W Walker, Heather Murray, Rosemary M McCloskey, P Richard Harrigan, and Richard H Liang. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. Plos One, 8(11):e78122, 2013.

[148] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. PloS One, 5(3):e9490, 2010.

[149] A Purvis. Using interspecies phylogenies to test macroevolutionary hypotheses, pages 153–168. Oxford University Press, 1996.

[150] Andy Purvis and Paul-Michael Agapow. Phylogeny imbalance: taxonomic level matters. Systematic biology, 51(6):844–854, 2002.

[151] Andy Purvis, Susanne A Fritz, Jesús Rodríguez, Paul H Harvey, and Richard Grenyer. The shape of mammalian phylogeny: patterns, processes and scales. Philosophical transactions of the royal society B: biological sciences, 366(1577):2462–2477, 2011.

[152] Andy Purvis, Aris Katzourakis, and Paul-Michael Agapow. Evaluating phylogenetic tree shape: two modifications to Fusco & Cronk's method. Journal of theoretical biology, 214(1):99–103, 2002.

[153] Yixuan Qiu, Jiali Mei, and Maintainer Yixuan Qiu. rARPACK: Solvers for large scale eigenvalue and SVD problems. R package version 0.12-0, 405, 2016.

[154] R development core team. R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria, 2008.

[155] Pasi Rastas. A general framework for local pairwise alignment statistics with gaps. In International workshop on algorithms in bioinformatics, pages 233–245. Springer, 2009.

[156] David F Robinson. Comparison of labeled trees with valency three. Journal of combinatorial theory, series B, 11(2):105–119, 1971.

[157] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. Mathematical biosciences, 53(1-2):131–147, 1981.

[158] Katy Robinson, Nick Fyson, Ted Cohen, Christophe Fraser, and Caroline Colijn. How the dynamics and structure of sexual contact networks shape pathogen phylogenies. PLoS computational biology, 9(6):e1003105, 2013.

[159] James S. Rogers. Response of Colless's tree imbalance to number of terminal taxa. Systematic biology, 42(1):102–105, 1993.

[160] James S Rogers. Central moments and probability distribution of colless's coefficient of tree imbalance. Evolution, 48(6):2026–2036, 1994.

[161] James S Rogers. Central moments and probability distributions of three measures of phylogenetic tree imbalance. Systematic biology, 45(1):99–110, 1996.

[162] Noah A Rosenberg. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Annals of combinatorics, 10(1):129–146, 2006.

[163] Michael J Sackin. "Good" and "Bad" phenograms. Systematic biology, 21(2):225–226, June 1972.

[164] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In Joint European conference on machine learning and knowledge discovery in databases, pages 313–325. Springer, 2008.

[165] Sangharsh Saini, Jagtar Singh, and Monika Saini. An improved branch-and-bound algorithm for cyclopeptide sequencing. Biometrics and bioinformatics, 6(6):154–158, 2014.

[166] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4):406–425, 1987.

[167] Emma Saulnier, Olivier Gascuel, and Samuel Alizon. Inferring epidemiological parameters from phylogenies using regression-abc: A comparative study. PLoS computational biology, 13(3):e1005416, 2017.

[168] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. Nucleic acids research, 40(D1):D13–D25, 2012.

[169] Klaus Peter Schliep. phangorn: phylogenetic analysis in R. Bioinformatics, 27(4):592–593, 2010.

[170] Charles Semple, Mike Steel, et al. Phylogenetics, volume 24. Oxford university press on demand, 2003.

[171] Oliver Serang. The probabilistic convolution tree: efficient exact Bayesian inference for faster LC-MS/MS protein inference. Plos One, 9(3):e91507, 2014.

[172] Jean Pierre Serre. A course in arithmetic, volume 7. Springer science & business media, 2012.

[173] Kwang Tsao Shao. Tree balance. Systematic zoology, 39(3):266–276, 1990.

[174] Wenhan Shao, Xinxin Li, Mohsan Ullah Goraya, Song Wang, and Ji-Long Chen. Evolution of influenza a virus by mutation and re-assortment. International journal of molecular sciences, 18(8):1650, 2017.

[175] Arthur Chun-Chieh Shih, Tzu-Chang Hsiao, Mei-Shang Ho, and Wen-Hsiung Li. Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. Proceedings of the national academy of sciences, 104(15):6283–6288, 2007.

[176] Joseph B Slowinski. Probabilities of n-trees under two models: a demonstration that asymmetrical interior nodes are not improbable. Systematic zoology, 39(1):89–94, 1990.

[177] Elliott Sober. Reconstructing the past: Parsimony, evolution, and inference. MIT press, 1991.

[178] Yun S. Song. Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees. Annals of combinatorics, 10(1):147–163, 2006.

[179] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth-death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). Proceedings of the national academy of sciences, 110(1):228–233, 2013.

[180] Ed Stam. Does imbalance in phylogenies reflect only bias? Evolution, 56(6):1292–1295, 2002.

[181] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9):1312–1313, 2014.

[182] Mike Steel. Distribution of the symmetric difference metric on phylogenetic trees. SIAM journal on discrete mathematics, 1(4):541–551, 1988.

[183] Mike Steel and Andy McKenzie. Properties of phylogenetic trees generated by yule-type speciation models. Mathematical biosciences, 170(1):91–112, 2001.

[184] Mike Steel and David Penny. Distributions of tree comparison metrics - some new results. Systematic biology, 42(2):126–141, 1993.

[185] M Stich and SC Manrubia. Topological properties of phylogenetic trees in evolutionary models. The european physical journal B, 70(4):583–592, 2009.

[186] Rahul Subramanian, Andrea L Graham, Bryan T Grenfell, and Nimalan Arinaminpathy. Universal or specific? a modeling-based comparison of broad-spectrum influenza vaccines against conventional, strain-matched vaccines. PLoS computational biology, 12(12):e1005204, 2016.

[187] Lajos Takács. A bernoulli excursion and its various applications. Advances in applied probability, 23(3):557–585, 1991.

[188] Minh Anh Thi Nguyen, Tanja Gesell, and Arndt von Haeseler. ImOSM: intermittent evolution and robustness of phylogenetic methods. Molecular biology and evolution, 29(2):663–673, 2012.

[189] Thu-Hien To, Matthieu Jung, Samantha Lycett, and Olivier Gascuel. Fast dating using least-squares criteria and algorithms. Systematic biology, 65(1):82–97, 2015.

[190] Luis Torgo. Data mining with R: learning with case studies. Chapman and Hall/CRC, 2011.

[191] Jana Trifinopoulos, Lam-Tung Nguyen, Arndt von Haeseler, and Bui Quang Minh. W-iq-tree: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic acids research, 44(W1):W232–W235, 2016.

[192] Yves Van de Peer and M Salemi. Phylogeny inference based on distance methods. The Phylogenetic Handbook, a Practical Approach to DNA and Protein Phylogeny, pages 101–136, 2003.

[193] Charles Van Loan. Computational frameworks for the fast Fourier transform, volume 10. Siam, 1992.

[194] Erik M Volz. Complex population dynamics and the coalescent under neutrality. Genetics, 190(1):187–201, 2012.

[195] Wei Wang and Choon Yik Tang. Distributed computation of classic and exponential closeness on tree graphs. In 2014 American control conference, pages 2090–2095. IEEE, 2014.

[196] Yao Wang and Xuelong Zhu. A fast algorithm for the Fourier transform over finite fields and its VLSI implementation. IEEE Journal on selected areas in communications, 6(3):572–577, 1988.

[197] Zhongde Wang. Fast algorithms for the discrete w transform and for the discrete fourier transform. IEEE Transactions on acoustics, speech, and signal processing, 32(4):803–816, 1984.

[198] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. Nature, 393(6684):440, 1998.

[199] Lucie Chan Mark J. Dominus Jin Ruan Rutger Vos William H. Piel, Jon Auman and Val Tannen. Treebase v. 2: A database of phylogenetic. E-BioSphere, 2009.

[200] Elizabeth Wolf, Joshua T Herbeck, Stephen Van Rompaey, Mari Kitahata, Katherine Thomas, Gregory Pepper, and Lisa Frenkel. Phylogenetic evidence of hiv-1 transmission between adult and adolescent men who have sex with men. AIDS research and human retroviruses, 33(4):318–322, 2017.

[201] Taoyang Wu and Kwok Pui Choi. On joint subtree distributions under two evolutionary models. Theoretical population biology, 108:13–23, 2016.

[202] Chris Wymant, Matthew Hall, Oliver Ratmann, David Bonsall, Tanya Golubchik, Mariateresa de Cesare, Astrid Gall, Marion Cornelissen, Christophe Fraser, The Maela Pneumococcal Collaboration STOP-HCV Consortium, and The BEEHIVE Collaboration. Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. Molecular biology and evolution, 35(3):719–733, 2017.

[203] Ziheng Yang. Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. Molecular biology and evolution, 10(6):1396–1401, 1993.

[204] George Udny Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC willis, FR S. Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character, 213(402-410):21–87, 1925.

[205] Xiaofan Zhou, Xing-Xing Shen, Chris Todd Hittinger, and Antonis Rokas. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. Molecular biology and evolution, 35(2):486–503, 2017.

Appendix A

Results Corresponding to Chapter 4

Linear combination of the classical tree shape statistics

Lemma. *The optimal λ value for the combination of a pair of statistics is always a real number.*

Proof. Consider linear combinations of the form $\lambda S_1 + S_2$, where S_1 and S_2 are vectors corresponding to two distinct real statistics. Without loss of generality we make two assumptions:

- S_1 and S_2 are orthogonal; this is because we can always write $S_2 = \alpha S_1 + S_3$, with $\alpha \in \mathbb{R}$ and S_3 orthogonal to S_1, and the resolution of $\lambda S_1 + S_2$ equals the resolution of $(\lambda + \alpha)S_1 + S_3$, in which the coefficient $\lambda + \alpha$ of S_1 is real if and only if λ is real, as $\alpha \in \mathbb{R}$.
- D_s is a real symmetric matrix (true for both resolutions).

Under these two assumptions, the aim is to find the value of λ that maximizes the resolution of $\lambda S_1 + S_2$ when S_1 and S_2 are orthogonal:

$$R_D(\lambda S_1 + S_2) = \frac{(\lambda S_1 + S_2)^t D_s (\lambda S_1 + S_2)}{(\lambda S_1 + S_2)^t (\lambda S_1 + S_2)} \tag{A.1}$$

Let us call the numerator and denominator of this equation f and g, respectively. The problem reduces to finding a λ such that the derivative of $\frac{f}{g}$ equals 0, which is equivalent to $f'g = g'f$ by the quotient rule. We thus have:

$$2[S_1^t D_s(\lambda S_1 + S_2)][(\lambda S_1 + S_2)^t(\lambda S_1 + S_2)] = 2[S_1^t(\lambda S_1 + S_2)](\lambda S_1 + S_2)^t D_s(\lambda S_1 + S_2)$$

After some algebra, the coefficient of λ^3 cancels out and we end up with the quadratic equation:

$$(S_1^t D_s S_2 S_1^t S_1)\lambda^2 + (S_2^t D_s S_2 S_1^t S_1 - S_1^t D_s S_1 S_2^t S_2)\lambda - (S_1^t D_s S_2 S_2^t S_2) := a\lambda^2 + b\lambda + c = 0$$

The discriminant $b^2 - 4ac$ of the quadratic function determines whether its roots are real. In this case, we note that $S_1^t S_1$ and $S_2^t S_2$ are non-negative real numbers, and we can easily see that the discriminant of the above equation is always non-negative, since the term $-4ac$ above is a perfect square and thus non-negative. Therefore, its roots are real, and so is λ. □

Tree shape Summaries Based on Network Science

Genbank accession numbers for outgroups

Virus	Outgroup	Genbank accession number
HIV-1 subtype B	a subtype D sequence	AY071949
Dengue virus serotype 4	isolate from the Philippines (1956)	U18433
Measles virus	a genotype D6 sequence	AY523581

Table A.1: Accession numbers for outgroups

Tree shape differentiates viruses and epidemiological scenarios

We find that tree shape carries considerable information and differs both between viruses and between different epidemiological scenarios for the same virus, on real as well as simulated data. While degree sequence, clustering coefficients and other measures based in network science are not informative for binary trees, a number of non-standard topological features differ. Furthermore, there are several features that distinguish well between groups of trees with the same overall level of asymmetry, highlighting the need to move beyond asymmetry when using trees to infer evolutionary and epidemiological parameters or to test hypotheses [68]. Figure A.1 and Figure A.2 illustrate the distributions of all tree statistics considered in chapter 4.

Most of the statistics do not vary much between the three viruses in Figure A.1 (red boxplots). Distinguishing the topologies in these groups of trees requires tools going beyond the traditional symmetry or imbalance metrics; in this case, the only statistically significant difference are produced by the number of cherries, weighted closeness centrality, maximum height, and the proportion of imbalanced subtrees (stairs) capture differences that are not apparent in the imbalance. In contrast, while two of the spectral statistics (maximum adjacency and maximum Laplacian eigenvalues) show statistically significant differences among the groups, the vertical scale is very small, and these differences are a small percentage of the overall value (2% for the maximum adjacency and 1% for the maximum Laplacian). Furthermore, three of the network statistics - namely, the Wiener index (meanpath) and

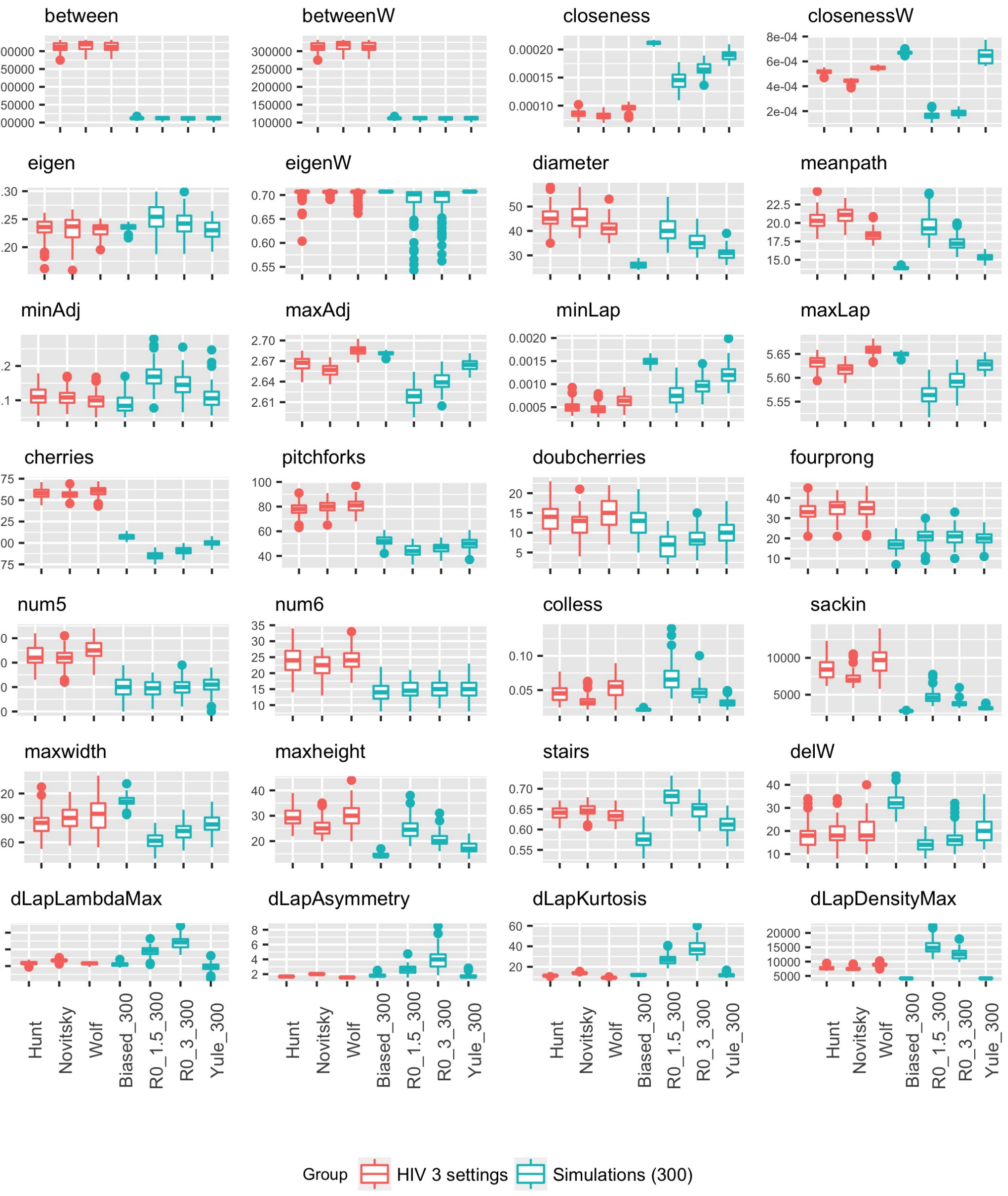


Figure A.1: Tree summaries for the HIV/Dengue/Measles, influenza and simulations of size 100.

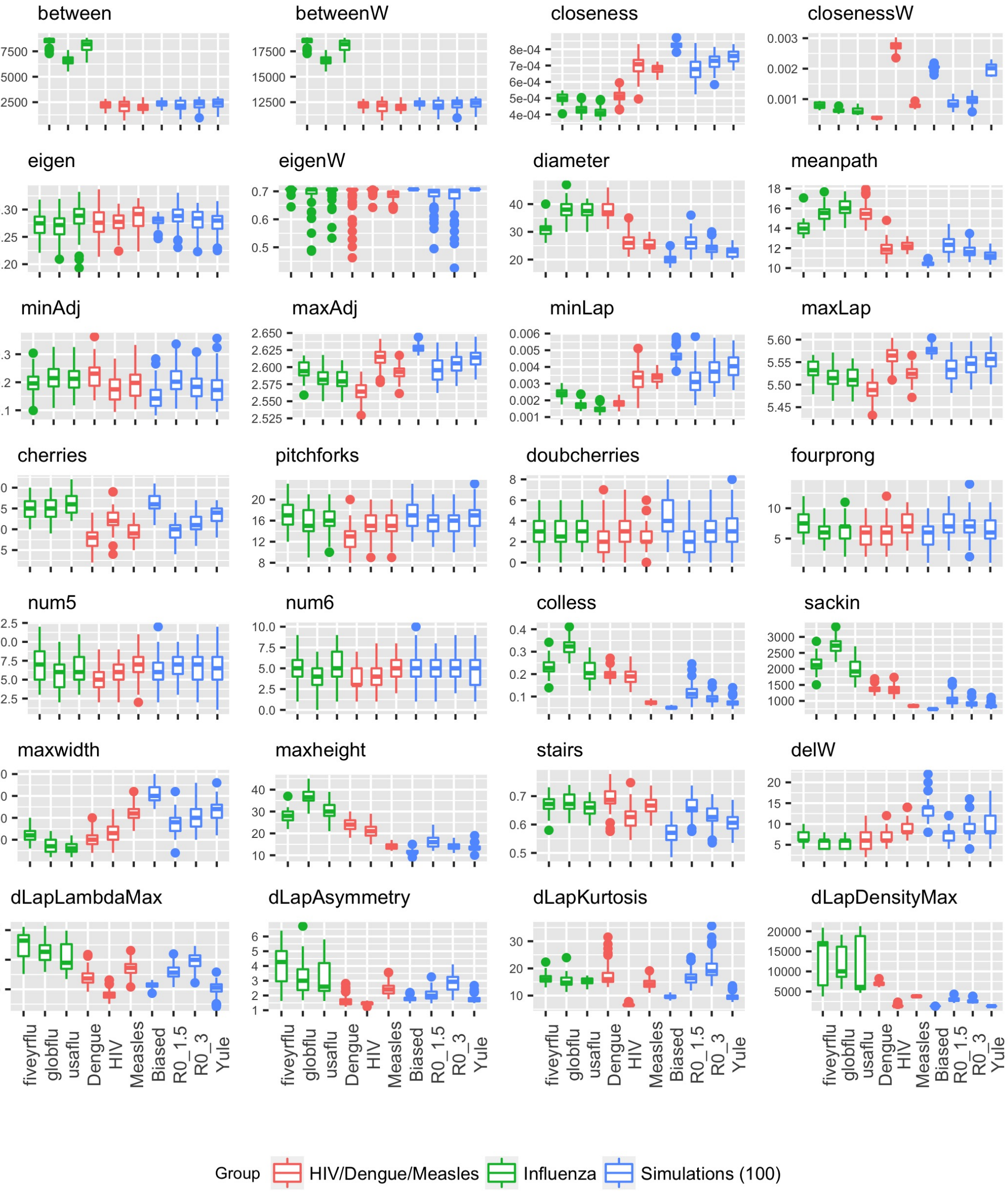


Figure A.2: Tree summaries for HIV in three settings and simulations of size 300

the weighted closeness and eigenvector centralities - also distinguish the three viruses, as do three of the four statistics based on the distance Laplacian.

The green boxplots in Figure A.1 show the tree summaries for influenza in three scenarios. It is well-known that global influenza patterns give rise to highly imbalanced trees, an observation which in part motivated the growing field of phylodynamics [68]. In our data, the global five-year and (2-year) USA flu trees have similar, lower, levels of asymmetry. The only statistics which are able to differentiate all three groups at a statistically significant level after multiple testing correction are betweenness centrality and the maximum Laplacian eigenvalue. As with the three viruses, the differences in the latter are not pronounced, and are very small in relative magnitude. In addition, the LM spectral statistics are quite discriminating for these scenarios, though none of them is able to do so in a statistically significant way. Figure A.3 shows a random tree from each group, and differences between the shapes of the flu trees are not immediately apparent by eye.

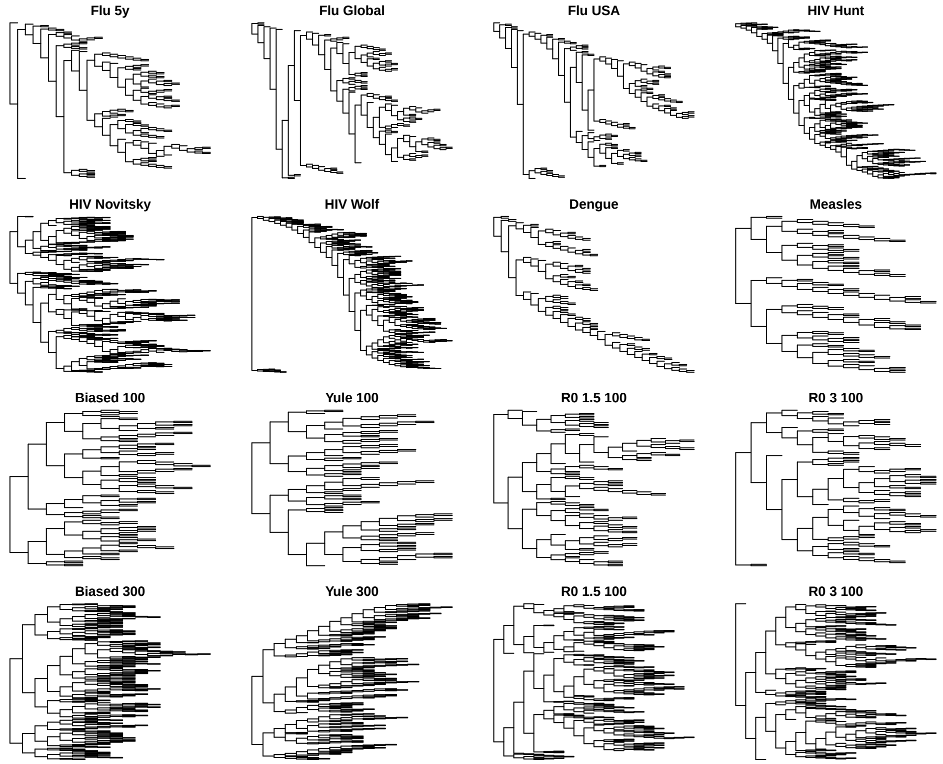


Figure A.3: A randomly sampled tree from each scenario (except HIV in the 3-virus comparison because HIV is represented in three other trees). To allow for focus on tree shape rather than on branch lengths, trees have been visualized with branch lengths set to 1.

The blue boxplots in Figure A.1 show the summaries for small trees (100 tips) in the four simulated settings. This time, only the closeness centrality, the diameter, and the mean path, as well as the number of cherries, the Sackin and Colless imbalances, the maximum height, the stairs statistic and the lambdaMax statistic based on the distance Laplacian spectrum, are able to differentiate every pair of these scenarios in a statistically significant way.

In contrast, the cyan boxplots in Figure A.2 show that large trees in the same simulated settings can be discriminated by all of these statistics, but also by several other ones, including the adjacency and Laplacian eigenvalues as well as the maximum width and the maximum difference in width (delW). This suggests that it is strictly easier to discriminate larger trees than it is to discriminate smaller trees.

Lastly, the three epidemiological scenarios (concentrated epidemic, generalized epidemic in a village, and generalized epidemic in a country) for HIV, shown in the red boxplots in Figure A.2, appear to be quite difficult to distinguish. Only the weighted and unweighted closeness, as well as the Sackin and Colless imbalance, the maximum eigenvalues of the adjacency and the Laplacian, and the asymmetry and kurtosis of the distance Laplacian spectrum, are able to do so in a statistically significant manner.

Appendix B

Results Corresponding to Chapter 5

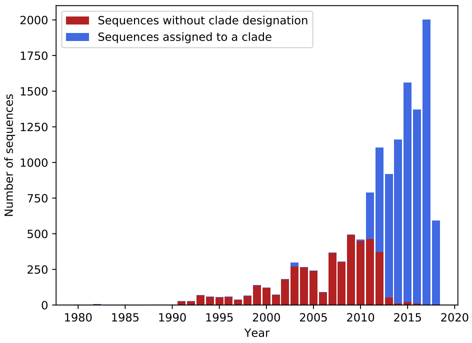


Figure B.1: Distribution of H3N2-HA sequences between 1980 and 2018. The blue bar corresponds to sequences assigned to one or more clades, whereas the red bar corresponds to sequences without clade designation. The rate of unassigned cases drops considerably for recent sequences (from 2011 on), which encompass most of the analyzed dataset (73.45%).

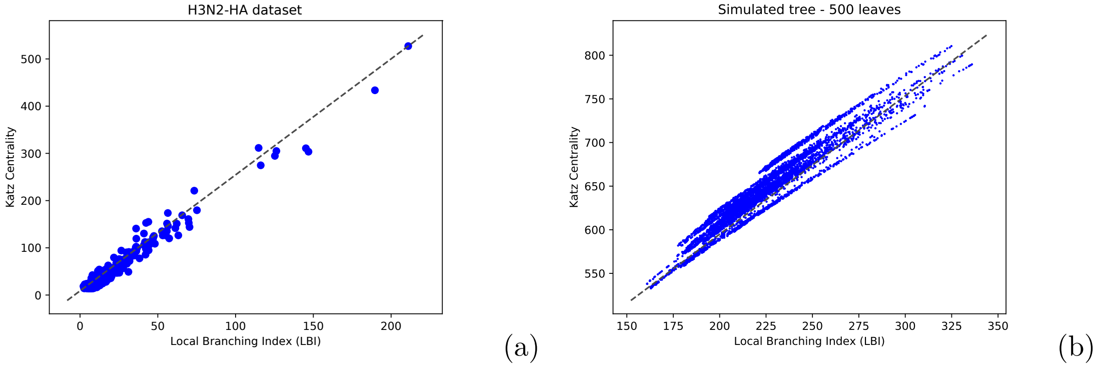


Figure B.2: Correlation between Local Branching Index (LBI) and Katz Centrality measures (weighted version, where the distance between two nodes is defined as the sum of the branch lengths in the path separating them). (a) Each point corresponds to the LBI and Katz Centrality measures computed for the root of the 391 subtrees from the H3N2-HA dataset. Parameters used: $\tau = 50$ (LBI), $\alpha = 0.95$ (Katz Centrality). (b) Correlation for 10 trees simulated with a pure birth process. All trees have 500 leaves, and each point corresponds to the LBI and Katz Centrality measures computed for a node of those trees. Parameters used: $\tau = 10$ (LBI), $\alpha = 0.95$ (Katz Centrality).