2019

Computational Gene Association Based on Relative Species Displacement in Gene and Morphology Trees

Martin D. Hellwig
University of Rhode Island, hellwig@uri.edu

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Recommended Citation

Hellwig, Martin D., "Computational Gene Association Based on Relative Species Displacement in Gene and Morphology Trees" (2019). *Open Access Dissertations*. Paper 846.
https://digitalcommons.uri.edu/oa_diss/846

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

COMPUTATIONAL GENE ASSOCIATION BASED ON RELATIVE SPECIES
DISPLACEMENT IN GENE AND MORPHOLOGY TREES

BY
MARTIN D. HELLWIG

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
COMPUTER SCIENCE

UNIVERSITY OF RHODE ISLAND

2019

DOCTOR OF PHILOSOPHY DISSERTATION
OF
MARTIN D. HELLWIG

APPROVED:

Dissertation Committee:

Major Professor Noah Daniels

Joan Peckham

Jonathan Puritz

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2019

ABSTRACT

While a number of methods exist for gene based fin prediction, these studies are mostly limited to single sites and to binary morphological characters and they are very susceptible to missing data. Here we present GATRS (Gene Association in TRee Space), a novel algorithm that can operate on continuous morphological observations such as fin size and is robust to missing data by operating in gene tree space, analyzing linkages between entire genes and morphological traits. GATRS performs a large number of comparisons between closely related trees which requires a tree distance method that is particularly apt at distinguishing closely related trees in a distant forest. Due to the number of comparisons the employed method's computational requirements must also be reasonable. Finally, gene trees are likely to contain substantial amounts of missing data to which any distance method deployed within GATRS must be reasonably robust. I conduct a thorough comparison of popular distance methods that are known to provide the best distance scores and find that none of the methods serve GATRS' purpose as the best performing methods are computationally prohibitive while the fastest methods falter in the presence of missing data. I therefore develop my own novel tree distance method called TDeft that focuses on taxon neighborhoods and produces scores of a quality rivaling that of the best methods while requiring modest computational time that only marginally exceeds that of the fastest methods. It also proves to be robust to a wide range of realistic scenarios. I test GATRS on a comprehensive dataset of osteichthyes with 211 taxa and 2072 loci and produce 160 implied associations. In spite of the sparse amount of existing research, I am able to demonstrate the correctness of more than half of the most significant associations, confirming the validity of my approach. I also report new discoveries for which no previous laboratory research exists and suggest that GATRS serve as

a guide for future biological research studies, reducing the need for expensive and invasive exploratory laboratory studies required to narrow the field of candidate genes.

ACKNOWLEDGMENTS

Compiling this dissertation would not have been possible without the support and assistance of my entire doctoral committee. I especially appreciate all the guidance I received from my major professor Dr. Noah Daniels. I would further like to extend my gratitude to the department chair Dr. Lisa DiPippo who has helped me overcome more than a few obstacles in the pursuit of this degree.

DEDICATION

To my parents without whom none of this would have been possible.

PREFACE

This dissertation is presented in manuscript format in accordance with University of Rhode Island Graduate School Guidelines. This dissertation is composed of three manuscripts that have been combined to satisfy the requirements of the department of Computer Science and Statistics. Chapter 1 introduces the problem and the particular challenges. Chapter 2 presents a comparison study evaluating the performance of popular tree distance computation methods. This study will be submitted to the journal *Systematic Biology*. In chapter 3 I develop a new tree distance measure that addresses many of the shortcomings of the existing methods and also provides additional data points used later on. This manuscript has already been submitted to the International Society for Computational Biology’s annual conference “Intelligent Systems for Molecular Biology (ISMB)”, which will be held in Basel in July of 2019. Finally, chapter 4 details the core component of this dissertation – an attempt to establish associations between genes and morphological traits in phylogenetic tree space. This work will be submitted to *PLoS Computational Biology*.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iv
DEDICATION	v
PREFACE	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	xi
LIST OF TABLES	xii
CHAPTER	
1 INTRODUCTION	1
1.1 Motivation and Significance	1
1.2 Summary of the Approach	2
2 PERFORMANCE OF EXISTING TREE DISTANCE MEASURES	6
2.1 Abstract	7
2.2 Introduction	7
2.3 Methodology	10
2.4 Results	12
2.4.1 Tree Sizes	12
2.4.2 Ties	13
2.4.3 Thresholds	14

	Page
2.4.4 Missing Data	16
2.4.5 Computational Requirements	18
2.5 Discussion	19
3 TAXON FOCUSED TREE DISTANCE ESTIMATION (TDEFT)	23
3.1 Introduction	24
3.1.1 Data Based Challenges	24
3.1.2 Analytic Challenges	27
3.2 Methodology	28
3.2.1 Neighborhood Analysis	28
3.2.2 Species Displacement	30
3.2.3 Species Displacement in Sets of Trees	31
3.2.4 Tree Distance	32
3.2.5 Formal Algorithm Description	32
3.2.6 Runtime for Larger Sets of Trees	33
3.2.7 Runtime for Pairs of Trees	34
3.2.8 Phylogenetic Networks	35
3.3 SIMULATIONS	35
3.3.1 Optimal Neighborhood Size	36
3.3.2 Comparison against other metrics on identical namespaces	40
3.3.3 Thresholding	43
3.3.4 Comparison against other metrics on heterogeneous namespaces	45
3.3.5 Processing Times	46

	Page
3.4 Results and Discussion	48
3.4.1 Future Directions	50
4 GENE ASSOCIATION IN TREE SPACE (GATRS)	51
4.1 Abstract	52
4.2 Introduction	52
4.2.1 Motivation and Significance	52
4.2.2 Existing Approaches	54
4.3 Methodology	57
4.3.1 Data Sets	57
4.3.2 Data Selection	57
4.3.3 Genetic Data	58
4.3.4 Morphometric Data	59
4.3.5 Missing Data	61
4.3.6 Alignment	62
4.3.7 Gene Tree Inference	62
4.3.8 Species Tree Inference	62
4.4 Gene Association in Tree Space (GATRS core)	63
4.4.1 Morphometric Space	63
4.4.2 Gene Space	64
4.4.3 Comparison between Morphometric and Gene Spaces	65
4.5 Results	69
4.5.1 Caudal Fin	70
4.5.2 Anal Fin	70

	Page
4.5.3 Dorsal Fin	71
4.5.4 Pelvic and Pectoral Fins	73
4.5.5 Pre-Orbital Length	73
4.6 Discussion	73
5 DISCUSSION	77
LIST OF REFERENCES	82

LIST OF FIGURES

Figure		Page
1	Performance on Trees of Different Sizes	13
2	Allowing Equality (left) or Requiring Strict Distance Increase (right)	14
3	Response to Thresholding	15
4	Response to Missing Data	16
5	Response to Missing Data with a 10% Threshold	18
6	n-away analysis 32 and 64 taxa at n=10	39
7	n-away analysis in a 1024 taxon tree at n=50	40
8	32-taxon 10-away tree analyzed using all 5 metrics	42
9	1024-taxon 10-away tree analyzed using the 4 fastest metrics. . .	42
10	Comparing metrics on a 64-taxon tree using a 0% threshold . .	44
11	Comparing metrics on a 64-taxon tree using a 10% threshold . .	44
12	Comparing metrics on a 64-taxon tree using a 20% threshold . .	44
13	Performance on a 64 taxon tree with 25% mismatched data . .	45
14	Response to Missing Data with a 10% Threshold	47
15	Distribution of c-values among statistically significant associations	67

LIST OF TABLES

Table		Page
1	Runtimes for 99 sequential comparisons (s)	19
2	Runtimes with $\tau = 4$	47
3	Caudal Fin Associations	70
4	Anal Fin Associations	71
5	Dorsal Fin Associations	72
6	Associations With Eye Position	74

CHAPTER 1

INTRODUCTION

1.1 Motivation and Significance

Morphological characteristics are heavily influenced by environmental pressures [1], epigenetic control [2] and, of course, information encoded at the nucleotide level [3]. In this study we attempt to detect signals that can associate given genes with morphological characteristics in spite of the significant complexity of the overall system and the large amount of noise that is introduced by factors such as environmental pressures and epigenetic control. As many gene regions control early developmental patterns, they often affect multiple morphological traits [4]. Similarly, each trait may be influenced by larger sets of loci. The relationship between genes and morphologies is hence a complex n-m relationship that cannot easily be modeled or explained. Determining which genes are responsible for particular morphologies can hence be very time consuming and expensive [5] as extensive laboratory experiments are necessary [6]. These experiments often involve animals, requiring even more resources and raising significant ethical concerns [7]. For example, in the widely used knockout study approach [8], researchers create mutant individuals by deactivating, deleting or mutating particular genes or loci in order to observe the effect on the organism [9]. This often produces animals with severe deformations or diseases that may be expected [10] or unexpected [11], which has led to a renewed discussion of the ethical tradeoff between animal welfare and research outcomes [12]. Other common studies aim to determine the effect certain external pressures have on organisms during their developmental stages. For example, certain fish populations in polluted waters display unusual and biomechanically disadvantageous [13] fin shapes that are not present in similar populations of the same species in cleaner waters [14]. This suggests that one or any combination

of the compounds present in the polluted environment act epigenetically on fish development and produce disadvantaged adults [15]. To determine which gene's expression levels are affected and to possibly link particular pollutants to their adverse effects, studies usually analyze gene expression in normal and abnormal individuals [16]. To achieve this, samples are treated with primary antibodies that specifically bind to the proteins coded by a particular gene. Secondary antibodies are used to visualize these primary antibodies using fluorescent properties [17]. As a result, the specimen can be visually analyzed to determine exactly where in a given tissue sample a certain gene is expressed. While very effective and commonly used, these studies are cost intensive as the primary antibodies must be created individually for each gene one wishes to visualize. Also, while a few genes can be visualized in the same sample, much larger sets of samples would be needed for an uninformed search for expressed genes. Such studies can be much more efficient if one has a general idea of which genes should or might be expressed in a tissue sample and which expressions one might want to visualize. By computationally determining which genes appear relevant to given morphologies, one can better focus experimental analyses. This, in turn, has the potential to reduce the number of needed laboratory studies as much of the preliminary and intermediate work can be replaced by computational analyses and simulations [18]. Better experimental focus obviously saves time and cost and can thereby increase the pace of overall discovery as well. In this work I will introduce and test a novel method to identify genes that are likely to have a direct or – more likely – indirect influence on certain morphologies. Further studies can then be focused on these candidate genes.

1.2 Summary of the Approach

Unlike traditional phenotype/genotype or genome wide association studies, my method does not operate on the molecular level. It instead aims to asso-

ciate nucleotide loci with morphologies or traits by working in tree space, which is naturally abstracted from the molecular level. I discuss the particular advantages of adding phylogenetic inference as a precursor to gene association in a later chapter. Generally, phylogenetic methods are subject to three types of polymorphisms: those carrying a phylogenetic signal, those representing a different and non-phylogenetic signal and, of course, noise. It is the aim of these methods to isolate the phylogenetic signal in order to establish likely evolutionary histories by filtering out noise and suppressing non-phylogenetic signals as much as possible [19]. This is an often challenging task as the non-phylogenetic signal can be stronger than the phylogenetic one and hence potentially yield misleading phylogenetic inferences [20]. While this non-phylogenetic signal is highly undesirable in phylogenetic inference, it can contain interesting information about the represented taxa. The study proposed herein will determine whether this includes a sufficiently strong signal to associate given genes with the presence and magnitude of morphological traits. GATRS (Gene Association in TRee Space) will compare the placement of taxa in the species tree with their placement in a set of gene trees. As a first step, all gene trees will be used to compute a consensus tree using ASTRAL II [21]. It is expected that there will be significant discord between the various gene trees. No individual tree can be expected to accurately reflect the phylogenetic history as all of them are subject to different evolutionary pressures. However, there has been general agreement that large numbers of trees inferred from different genes on the same species set can be used to infer a species tree that will reflect the actual evolutionary history with reasonable accuracy [22]. While in recent years there has been some discussion about the validity of coalescent consensus methods for species tree estimation, the most recent studies strongly suggest that they produce valid and reliable estimates – even in the presence of incomplete

data [23]. Once all gene trees and the consensus tree have been created, GATRS will iterate over the entire set of taxa in the consensus tree. It will then determine how similar each taxon’s placement in each of the gene trees is when compared to the consensus tree. In doing so, GATRS is able to detect particular genes of interest for each taxon as a divergent displacement in a gene tree might carry a non-phylogenetic signal that may later be used to associate it with a particular trait. It is obvious that each trait will be influenced by a set of loci and that their impacts will vary greatly. As a result, trees built on less impactful genes will more closely resemble the species tree while more influential genes will yield trees that contain more discord. It is important to note that even in such a discordant tree, many taxa will be placed in very similar neighborhoods as they may not display any trait controlled by the underlying gene. Others, however, should be grouped closer to taxa having similar trait expressions and further from their phylogenetically correct neighbors. Hence the magnitude of displacement in each gene tree will be compared against the magnitude of each taxon’s particular morphological characters. Genes’ associations will be computed based on the overall correlation between these two data streams across all taxa. A particular challenge that arose during the development of GATRS was the choice of available tree distance measures. As GATRS aims to associate minor incremental changes in trees and a given taxon’s position within them, distances have to be computed with a method that is able to accurately report such incremental changes. In chapter 2 of this dissertation I present a comparison of the most popular tree distance measures and their performance under realistic scenarios of missing data. This is important as GATRS will operate on highly incomplete data and any employed tree distance metric must hence be somewhat robust to missing data. During this comparison study it became obvious that some of the existing approaches performed reasonably well but

required unreasonable amount of computational resources while the faster methods proved to be inadequate when operating on datasets with larger shares of missing data. I therefore developed a new tree distance calculation method called TDeft which I introduce in detail in chapter 3. TDeft provides a reasonably balanced performance portfolio. It is nearly as fast as the fastest existing methods and in most cases produces score qualities that are close to the best competing methods. TDeft also produces individual taxon displacement scores which GATRS uses for its displacement calculations. Chapter 4 of this dissertation introduces GATRS in detail and reports on its performance using a large biological dataset.

CHAPTER 2

PERFORMANCE OF EXISTING TREE DISTANCE MEASURES

This manuscript is being prepared for publication in the journal Systematic Biology

2.1 Abstract

Phylogenetic analyses often require comparisons between sets of trees, frequently comparing them to a benchmark tree. It is particularly important that a distance method be able to accurately differentiate between two similar trees in a forest of trees that are all very different from the benchmark tree as this allows for detection of incremental changes. Many tree distance methods exist and their performance varies greatly in different scenarios. Here we compare a number of popular distance methods by building on an existing comparison study and addressing certain shortcomings within it. We report the methods that are fastest as well as most robust to noise under varying degrees of missing data and suggest specific methods for a variety of studies and scenarios. Given unlimited computational resources, Align proved to be the method of choice except in the presence of large amounts of missing data. In those cases MAST generally proved more robust. However, both methods are computationally expensive and so the fastest method, Robinson Foulds emerged as the preferable method on large datasets or within the confines of limited computational resources. However, due to its tendency to plateau very early, Robinson Foulds could not be recommended for studies involving inconsistent taxon namespaces. In the course of this study it became quite obvious that in order to choose the best tree distance method, one needs to have a good understanding of the nature of the available data and that there is room for a new algorithm to provide a reasonable compromise between computational requirements and scoring quality.

2.2 Introduction

The field of phylogenetics is primarily concerned with estimating evolutionary histories and relationships between extant taxa. These histories are most commonly modeled as bifurcating trees with each leaf node representing a particular

taxon. There is no shortage of competing methods to infer such trees from nucleotide sequences [24], but in order to compare the various estimations - perhaps against a known ground truth - one needs to rely on tree distance measures that compare two or more trees against each other and provide an estimate on how different trees are or how far apart they would be in a metatree. A handful of popular methods exist and they are mostly divided into two categories - those that use branch lengths and those that do not [25]. Some of the former use branch lengths exclusively and some pair them with structural measurements while the latter use purely structural approaches. It is not uncommon for competing methods to yield different - often contradictory - results and it is therefore important to know which method might perform best under given conditions.

In 2015, Kuhner and Yamato published a comprehensive study comparing various distance metrics and they came to the conclusion that structural methods yield better results in almost all instances [26]. However, even within the structural methods there were large differences in the quality of computed distance scores and ultimately a branch alignment method referred to as Align [27] proved to consistently perform best. Our own research confirmed this, but it also showed that this method is by far the slowest and quickly becomes infeasible in a realm of limited computational resources. One of the weaknesses of Kuhner and Yamato's otherwise robust study is the fact that they do not sufficiently consider runtime of the various algorithms. Here we present a study that compares those methods reported by Kuhner and Yamato to yield the best distance scores in terms of both quality and computational requirements. In order to ensure easy comparability, we will use the same nomenclature introduced by Kuhner and Yamato and refer to the various methods by the same names.

We will also use a similar incremental change analysis - or n-away analysis -

as the one found in their study. We will, however, address a second limitation by treating ties differently than they did. In an n -away study, we consider a given base tree and compare it against two similar trees in a distant forest. One of these trees will be n steps removed from the base tree, the other will be $n+k$ steps removed from the same base. Naturally the distance between the two trees is k steps. As n grows larger, differentiating trees that are only very small k removed from each other becomes increasingly harder. However, real world applications often require exactly this ability to differentiate between trees in such a distant forest in order to detect incremental change. Therefore a method’s ability to accurately report such close trees as different is an excellent indicator of its quality. Essentially we require a method to provide a higher distance score for an $n+k$ -away tree than an n -away tree in order to consider the classification successful. This becomes problematic in the presence of ties and several methods tend to provide frequent ties at larger n . We are then faced with the decision whether to count this as a proper classification or as a failure to classify the trees as different. Kuhner and Yamato chose to count ties as half successes and while this is certainly a valid approach, we believe that it would be better to avoid the tie problem altogether by changing the classifier. Instead of requiring a method to yield a higher score for an $n+k$ -away tree, we require it to exceed a certain threshold ϵ . We define this as a percentage of the lower distance score, so in order to classify a differentiation as successful we require that $f(n+k) \geq f(n) * \epsilon$.

Finally, we improve Kuhner and Yamato’s analysis by considering situations of incomplete taxon namespaces due to missing data. In their 2015 study, the authors compare trees containing the same taxa in different confirmations. This provides a somewhat unrealistic test case since in real world scenarios, data are rarely complete and certain taxa are bound to be present in some trees and ab-

sent in others. This poses a problem since some tree distance metrics - notably Robinson Foulds [28] - tend to classify trees with inconsistent namespaces as quite different and are less able to differentiate between trees in distant forests due to their tendency to plateau rather early [29]. In this study we compare the measures under a set of different amounts of missing data and show which method might be most appropriate in what scenario. Our results may serve as a guide to selecting the best available method for future research studies as there is clearly no one superior method that can be used indiscriminantly for all studies involving all types of data.

2.3 Methodology

We begin by creating test scenarios for various tree sizes and amounts of missing data. Each scenario is built on a base tree and a certain number of altered trees. We construct an altered tree by using the popular subtree prune and re-graft method (SPR) [30]. This method selects a bifurcating node at random, removes one of its children with any existing descendants and inserts it at a second random location of the tree by creating a new bifurcation along an existing edge. The original parent of the dislocated subtree is eliminated as a node and replaced by a straight edge. We call this a 1-away tree and continue to produce 2-away, 3-away and ultimately n -away trees by performing additional rounds of SPR on the modified trees. In this study we produce trees up to $n=100$. This yields a single set of trees and a single test case but in order to compute the likelihood of a method to correctly differentiate between n -away and $n+k$ -away trees, we need to create multiple such test scenarios. In this study we built 1000 independent sets of base trees and their respective 100 n -away trees. This produced a total of 100,000 trees for each tree size and in this study we considered six different tree sizes: 32 taxa, 64 taxa, 128 taxa, 256 taxa, 512 taxa and 1024 taxa. For each

tree size we also create alternate scenarios with inconsistent namespaces to test a method’s robustness to missing data. We remove 10%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 75%, 80% and 90% of taxa respectively.

For each tree size and each level of missing data we then run an incremental change analysis for the tested methods on 1000 replicates. For each replicate we initially compute a distance score between the base tree and the 1-away tree. We then compute a distance score between the same base tree and the corresponding 2-away tree and compare the two obtained distance scores. Subsequently we determine whether the distance method properly differentiated the 1-away and 2-away trees by analyzing whether the 2-away score is larger than the 1-away score multiplied by a given threshold. This threshold will be constant throughout the entire comparison across replicates, distance measures, tree sizes and missing data scenarios. In a later step we rerun the entire analysis using different thresholds to point out how robust the different methods are to such a threshold and, consequently, to noise in tree confirmations. The described step is an $n+k$ -away analysis with $n=1$ and $k=1$. We repeat this step for all k until we exhaust the simulated tree space at $n=1$ and $k=99$. We then continue by incrementing n and continue with $n=2$ and $k=1$, again computing every comparison until we reach $n=2$ and $k=98$. Eventually we arrive at $n=99$ and $k=1$. For each of these 5,000 comparisons we decide whether the method accurately differentiated between the n -away and the $n+k$ -away tree. This yields a binary result matrix for one replicate. By repeating this process for the remaining 9999 replicates, we create a binary result cube that we compress into a final result matrix by computing the percentage of replicates in which a method proved successful at a given n and k . This matrix is unique to each method, tree size and amount of missing

data. We can then use these matrices to plot the various methods' performances against each other. Theoretically we would expect the resulting success curves to be asymptotic, but in spite of the rather large set of 1,000 replicates some noise must be expected. To reduce minor fluctuations in the reported graphs, we apply a conservative smoothing algorithm that adjusts each observation to the mean of itself as well as its two nearest neighbors. At very large $n+k$ and comparably small tree sizes, we would also expect correct classification rates to begin declining as the theoretical space of possible tree confirmations becomes exhausted and additional mutations are increasingly likely to resemble a previously analyzed permutation.

2.4 Results

2.4.1 Tree Sizes

It should be noted that accurate classification of neighboring trees in distant forests becomes easier as the size of the tree increases. The larger the tree, the more obvious variations will be and the smaller it is, the sooner the space of possible confirmations is exhausted. This becomes obvious when we consider a four taxon tree. Such a tree can be built in exactly three different ways and so once we applied an initial change, there are only two other options left for a subsequent change. This means that the second change has a 1/3 chance of producing the original tree instead of a new different one. Even if we manage to provide such a novel tree in our second and third manipulations, the fourth is guaranteed to produce a tree structure that has previously been used. That structure is obviously closer to the original base tree than the previous one, making $f(n+3) < f(n+2)$. The same limitation exists in larger trees although the theoretical conformational space grows rapidly with tree size. Figure 1

illustrates the drastically improved classification rates at the same distance n and the same threshold for 32, 64 and 128 taxon trees respectively. We will explain the importance of thresholds in a later section.

Figure 1: Performance on Trees of Different Sizes

2.4.2 Ties

The core premise of this work states that in order to accurately differentiate between n -away and $n+k$ -away trees, a distance method has to produce a considerably higher score for the $n+k$ -away tree. Many methods produce ties between neighboring trees even though they are actually different from each other. This effect is particularly noticeable among the purely topological methods [26]. Kuhner and Yamato chose to consider such ties as half successes which is likely reasonable in certain scenarios. Kuhner and Yamato state that treating ties as failures would put purely topological methods at a disadvantage. We were not able to confirm this supposed effect in our study although it may likely exist in other settings. As we are only comparing topological methods herein, we need not worry about skewing the relative performance between topological methods and such that take distances into account. Furthermore, we argue that a tie is a fundamentally different result than a ranking of two separate trees. When trees are not identical, methods that provide a ranking more often are strictly preferable over those that more commonly yield ties. Therefore we do not count ties as

successes or partial successes but classify them as failures. This seemingly minute change produces drastically different success ratios for some of the tested methods while others show little or no change. As a result, the relative performance gain between the various tested methods changes considerably as the example of a 64 taxon tree at $n=30$ illustrates (Figure 2). At this distance, counting ties as half successes yields a large reported boost in performance for MAST [31] and Robinson Foulds [28] while the classification plots for Node [32] and Align remain identical. As a result, one might falsely consider Robinson Foulds to be equivalent or superior to Align [27]. Yet when counting ties as failures, the measure becomes inferior. This speaks to its limited robustness to noise in tree space as it is not as easily able to distinguish between closely related trees.

Figure 2: Allowing Equality (left) or Requiring Strict Distance Increase (right)

2.4.3 Thresholds

Since the rather minute change of counting ties as half successes can noticeably change the relative performance of different methods, it is only reasonable to also examine situations in which very small differences in distance scores are reported. One has to expect a certain amount of noise in tree space and so trees that are very similar may accidentally be scored as either identical or even different in the opposite direction. Similarly, identical trees might accidentally be scored

as minutely different. We will eliminate this risk by imposing a certain threshold and classifying any reported difference within that threshold as a tie and hence as a failure to differentiate. Only if the $n+k$ -away tree is reported as being further removed than the n -away tree by some given threshold, will we count this classification as successful. Even when we require just a 1% threshold, the relative performances change significantly (compare results in figure 3) and at a 5% threshold this trend becomes even more obvious.

Thresholding may lead to an accurate result as being classified as failure if the two compared trees are indeed close enough to be removed from each other by less than the defined threshold. In this case a distance method might accurately report them as being very close yet its results would be interpreted as a tie and hence as a failure to differentiate between the two trees. However, every method would be affected by this adverse classification equally, maintaining the validity of our study as we compare the various distance methods' performance. It should also be obvious that only a relatively small number of neighboring trees fall within the defined threshold while every single comparison between two neighboring trees benefits from the noise reduction effect thresholding produces. The benefits hence clearly outweigh the adverse effects.

Figure 3: Response to Thresholding

Figure 4: Response to Missing Data

2.4.4 Missing Data

When working with biological data, it is very common to encounter inconsistent taxon namespaces. Any two compared trees will consist of two separate sets of taxa which may have a limited intersection. As an immediate result, the two trees will also only have a certain number of shared edges. Comparing trees be-

comes harder with smaller namespace intersections since, for example, a 64 taxon tree with 50% missing taxa will – at best – yield as much structural evidence as a 32 taxon tree. Therefore, correct classification rates will decline in the presence of missing data as figure 4 illustrates on the example of a 128 taxon tree with a modest threshold requirement of 5% at a distance of $n=10$. It should be noted that we only show correct classification rates above 50%. Figure 4 (a) shows the classification rates without missing data. Unsurprisingly, Align performs best while the other methods respond marginally slower. Clearly there is no particularly noteworthy advantage to either method in this case and one would likely select an approach based on other criteria such as computational complexity.

When we consider two trees in which ten percent of the taxa are mismatched, we can notice a slight shift in relative performance. MAST appears to perform slightly better in this situation, essentially approximating Align’s outcome. Robinson Foulds shows a very slightly lower classification rate. Both of these changes are minor in the 10% scenario, but as we remove more taxa, we can detect a trend. Robinson Foulds becomes less effective at classifying similar trees in distant forests and once we exceed 50% of missing data, it is entirely incapable of classifying trees with better than random accuracy throughout our observation window. This is not surprising since this metric is known to consider similar trees with limited taxon intersections as different and to plateau quite early [29]. Node also yields declining classification successes with increased missing data rates, although their decline is slower than that observed in Robinson Foulds. Interestingly, the commonly superior method Align is increasingly outperformed by MAST as more taxa are mismatched between the compared trees. While Align’s correct classification rates remain very high, but MAST is clearly more robust to this scenario. This effect becomes amplified when we impose a threshold of 10% as shown in figure 5.

Figure 5: Response to Missing Data with a 10% Threshold

2.4.5 Computational Requirements

While accuracy and robustness to noise and missing data are highly relevant, any practical method must also be usable in a normal research setting with limited computational infrastructure. Hence the runtime and overall computational requirements are equally important features to consider when selecting a distance method. Obviously the actual runtime depends on a variety of factors such as the deployed architecture and any other processes that may be competing for resources. As many of these can be background processes, it is near impossible to give pre-

cise forecasts for the runtime of any method. However, when tested on the same architecture and under the same external conditions, the reported experimental runtimes serve as an excellent guide in comparing the methods against each other and in selecting whichever approach might be best suited for any given purpose. Here we used an 18 core CPU that presents 36 virtual cores through the process of hyperthreading. Hyperthreading allows a physical CPU core to use existing slack in one process to process another one simultaneously [33]. This, however, can only work when one of the two processes actually has significant amounts of slack [34]. When both processes run very efficient algorithms, it is likely that one of them will have to wait for the CPU core to become available while the other one is processed exclusively. Our tests were performed on a Microsoft Windows 10 platform which, due to its high level of hardware abstraction, does not allow us to select a specific CPU core for a particular thread. It is hence expected that running the exact same test multiple times will yield substantially different processing times [35]. We counteract this problem by taking the mean processing time of 200 iterations of each test run. Table 1 reports the mean processing time required to compute 99 independent comparisons between trees of the given size, each replicated 200 times.

Table 1: Runtimes for 99 sequential comparisons (s)

Method	32 Taxa	64 Taxa	128 Taxa	256 Taxa
Align	3.175 ($\sigma=0.52$)	20.79 ($\sigma=1.93$)	189 ($\sigma=12.1$)	3,378 ($\sigma=58.7$)
MAST	3.7 ($\sigma=0.43$)	12.44 ($\sigma=1.19$)	47.87 ($\sigma=3.16$)	178 ($\sigma=7.4$)
Node	0.7 ($\sigma=0.24$)	2.38 ($\sigma=0.37$)	10.08 ($\sigma=0.77$)	87.45 ($\sigma=4.61$)
Robinson Foulds	0.4 ($\sigma=0.18$)	0.65 ($\sigma=0.21$)	1 ($\sigma=0.38$)	1.42 ($\sigma=0.43$)

2.5 Discussion

Selecting the appropriate tree distance method has important implications for the quality and the feasibility of a proposed study. To make such a decision,

researchers need to know how well the different options perform under certain scenarios. While existing research provides a solid comparison of popular distance calculation methods’ performance in an incremental change analysis, our comparison adds score thresholding to test their robustness to noise. We also comprehensively test the method’s performance when analyzing trees with limited taxon intersections. Our results show that there is no one superior method, but that one has to carefully consider the nature of the data one is working with as well as the computational resources that may be available. Especially when performing large numbers of comparisons or when working with large trees, expected runtime becomes a major consideration. In our analysis, a single comparison between two 256 taxon trees could be performed in as little as 0.015 seconds using Robinson Foulds or as much as 34.12 seconds when choosing Align. Note that these figures are averages for a single comparison while table 1 lists average runtimes for 99 sequential comparisons. These times were both produced by single 2.5GHz cores which makes them reasonable expected runtimes for an average research project without dedicated architecture. However, as our results also show, runtimes increase at very different paces among the compared methods. Align, for example, grows roughly quadratically which is in line with the theoretical foundations of the algorithm [27]. As a result, a single comparison of two 2,048 taxon trees would already consume 36 minutes. Since many studies will require large numbers of such comparisons, this may quickly become prohibitive. However, with relatively complete data, Align yields by far the best results, reliably differentiating between relatively close trees in a distant forest. It maintains this superior performance across different tree sizes. When we impose a threshold to test the methods’ robustness to noise and the strength of their reported scores, Align increases its performance gain over the competing methods. Given unlimited computational resources it should hence

be the method of choice on near-complete sets of data. Align’s major weakness became evident when we tested it on trees with incomplete data or heterogeneous taxon namespaces. In the absence of large taxon intersections between trees, Align quickly lost its superiority and yielded increasingly disappointing results. It still outperformed most other metrics, but considering its much higher computational requirements, this relative gain in distance score quality makes it a less desirable choice in the presence of missing data. Interestingly, MAST proved to be exceptionally robust to this test and while it was typically outperformed by most other methods on near-complete sets of data, it proved to be clearly superior when operating on small taxon intersections, even improving its relative advantage when we imposed a scoring threshold. MAST was the second slowest method in our tests, but given sufficient resources, it would be an excellent choice when operating on incomplete data. Node yielded reasonably good results but was almost always outperformed by other methods in our tests. It presents a midfield solution that does not appear to justify its midfield computational requirements since it was often outperformed by the faster methods. There may well be specialized applications for Node as it appears to be a valid distance method with reasonable performance, but in our n-away tests we found no scenario in which we would recommend it over the competing methods. Robinson Foulds’ most impressive advantage are its extremely limited computational requirements. It consistently proved to be the fastest method by far, processing a single comparison on a 256 taxon tree in a matter of 15 milliseconds. Robinson Foulds also yields excellent scores for near-complete datasets. However, it shows a major weakness in the presence of incomplete taxon namespaces across trees. When comparing trees with limited taxon intersections, Robinson Foulds is easily outperformed by all other measures and quickly drops below a 50% correct classification rate throughout the obser-

vation window once we impose a score threshold. It still presents an excellent choice for studies with very good data and limited computational resources, but one should exercise caution when using Robinson Foulds on inconsistent datasets. When choosing a distance calculation method, one has to carefully analyze the set of data one is working with and particularly determine the amount of missing data that may be present. This will determine which distance method can be expected to yield the most beneficial distance scores as illustrated in this study. As a second consideration, one has to consider the candidate methods' computational requirements and balance their complexity against the available infrastructure and potentially determine a reasonable compromise between speed and scoring quality. This decision has to be made individually for every study and the decision process greatly benefits from comparison studies that test different methods under realistic scenarios. In this study we have built on an existing comparison study and augmented it with additional testing criteria to facilitate the selection of an appropriate distance method for GATRS and a wide range of other future studies. Unfortunately we have found that no one method provides a sufficiently reasonable balance between runtime and scoring quality that would be suitable for GATRS and we will hence develop an entirely new distance method to suit our particular needs.

CHAPTER 3

TAXON FOCUSED TREE DISTANCE ESTIMATION (TDEFT)

This manuscript has been submitted to the International Society for Computational Biology’s annual conference “Intelligent Systems for Molecular Biology (ISMB)”, which will be held in Basel in July of 2019.

3.1 Introduction

The field of phylogenetics is primarily concerned with inferring ancestral histories based on current observations. Many algorithms exist to construct and align nucleotide sequences and to use those alignments as bases for phylogenetic trees. Typically these trees are built on an individual locus, although in some applications trees are built using concatenated sequences. In this case, the sequences for multiple loci are concatenated and aligned into an artificial concatenation that does not as such exist in the original nucleotide sequence. The concatenated loci may or may not be adjacent in the original genome and they need not be in the same order, although order must obviously be maintained within the concatenated alignment of different samples. Concatenation based trees are a popular method of inferring evolutionary histories although in recent years the consensus tree approach has become more prominent as the concatenation approach is statistically inconsistent [36]. In a consensus tree approach sets of trees are built on individual loci and merged into a single consensus tree that approximates the underlying evolutionary history. Inferring such a species tree is often the primary focus of phylogenetic studies, but knowing how different two trees are from each other can produce important clues about the underlying data or it can simply be used to test and validate tree construction methods. A number of existing algorithms are available to compare trees and determine their distance from each other, but most of them are limited by one or more of the challenges that all comparison algorithms face.

3.1.1 Data Based Challenges

Some tree distance methods partially or fully rely on the path lengths between pairs of taxa [25]. This approach has been successfully applied in many scenarios, but it is bound to fail in the absence of branch length information. Lack of branch

lengths is therefore a significant challenge for tree distance estimation as some of the best measures are unable to process such trees. Branch lengths could be absent for a variety of reasons but is particularly common when operating on consensus trees such as species trees. Most consensus tree methods such as the popular and powerful ASTRAL II [21] can only provide support values for edges, but not branch lengths [37]. When operating purely on gene trees or concatenation trees, edge lengths are available in most cases as these trees aim at describing evolutionary processes [38].

A second significant challenge posed by phylogenetic datasets is the inherent inconsistency of taxa in trees built on different loci. It is very common for some considered species to be fully sequenced while other species have either not been sufficiently sequenced or their genome is insufficiently annotated. As sequencing cost have drastically declined in recent years [39], the scientific community has amassed large collections of reads, but even if a particular locus for a particular species has been sequenced and is available in a public database, it can only be used as the basis for a gene tree once it has been properly identified and annotated as belonging to that gene [40]. Hence at least some of the genes of interest are usually unavailable, even for well sequenced taxa. This produces a significant amount of missing data and means that many of the analyzed species will be present in some trees and absent from others. As a result, namespaces - representing the set of taxon names present in a given tree - are most commonly inconsistent across even small sets of trees. However, many topology based methods require trees to have identical namespaces [41]. One common approach to correct for this diversity involves trimming trees to only include those taxa that are shared by all of the compared trees. Tree edit distance methods [28] perform this task implicitly. In limited datasets this can actually improve accuracy [42]. However, when comparing

multiple trees that are based on real world sequencing data, it is very common to have only small intersections of taxa across all trees. It is, therefore, often impossible to produce a valid matrix of tree distances on larger datasets using such an approach. Our algorithm focuses on those species that are present in just one tree at a time and it is thereby capable of operating on pairs of trees with relatively small taxon intersections. As TDeft only considers one pair of trees at a time, its performance is not affected by the size of taxon intersection with the remaining trees in the global set of trees. In this fashion it can operate effectively on such larger sets that may not have a joint intersection at all.

Like most other phylogenetic operations, tree distance calculations are also computationally expensive, requiring algorithms that balance the quality of reported results with computational requirements. Longer runtimes have long been acceptable for tree comparison methods as datasets only grew slowly and the underlying trees needed considerable processing time themselves. However, tree distance metrics have not kept up with the rapidly increasing speed of discovery and sequencing [43] as well as faster alignment [44], phylogenetic tree inferring [24] methods and consensus tree estimation algorithms [45]. And while there have been some gains in hardware performance [46], the increased processing power is vastly outpaced by the growing size of available data [47]. This is still not particularly problematic for traditional studies that only rely on a limited number of comparisons. More novel approaches, however, require a much larger number of interarboreal comparisons. It is obvious that in the presence of near-constant processing power and quickly growing datasets, only a much more efficient algorithm will enable these future studies.

3.1.2 Analytic Challenges

The aforementioned challenges result from the nature of the available data and, of course, the computational complexity of the tree distance problem itself. On the output side, one of the hardest challenges for tree distance methods requires accurately differentiating between trees in a distant forest. When observed from a large distance, closely related pairs of trees may falsely appear identical. Detecting and reporting the incremental change between trees that are already very different from some original base tree is a difficult problem and hence incremental recombination analysis has evolved as a leading approach to test tree distance methods. Also known as n-away test [26], this method compares an initial tree, or base tree, with at least two other trees. Distances will be computed between the base tree and each of the remaining trees and a method’s classification is considered successful when it accurately distinguishes between the two trees purely on the basis of the computed distances from the base tree. Some measures such as Robinson Foulds are known to perform relatively poorly in this test as they plateau quite early [29]. Accurate classification of such distant and similar trees is obviously very important for any study focusing on incremental change analysis. We will introduce and analyze the sensitivity of distance methods in a later chapter.

While developing GATRS, we found the need for a tree distance metric that could address several of these major challenges simultaneously. We particularly needed an algorithm that would allow for extensive comparisons between heterogeneous evolutionary trees of both rooted and unrooted varieties on limited computational architecture. The nature of our study further required a taxon-focused method that could report differences in relationships between taxa. Kuhner and Yamato provide a solid comparison of numerous existing tree distance metrics [26] and we mostly considered the methods that performed best in their study. To

maintain comparability we adopt their nomenclature, referring to the same methods by the same names they defined. However, we found that none of the popular distance methods were able to serve our purpose.

Kuhner and Yamato investigated performance differences between topological methods as well as distance methods and those that combine the two approaches. As alluded to earlier, branch lengths may not always be available for our purposes which will limit us to the purely topological methods such as like Trip [48], MAST [31], Align [27] and the Robinson Foulds algorithm [28]. The latter tends to plateau too soon and can hence not reasonably identify incremental changes [29].

3.2 Methodology

3.2.1 Neighborhood Analysis

As previously stated, our analysis is mostly concerned with relationships between closely related species and the differences that exist in these relationships across sets of heterogeneous trees. We can therefore assume a taxon-centric focus on the taxon and its neighborhood within any given tree rather than a tree-centric focus that primarily accounts for changes in the larger structure or global scale distribution of taxa within it. We begin by defining a local neighborhood for each taxon and concentrate our further calculations on the related taxa that inhabit the same neighborhood. There are numerous ways of defining neighborhoods [49] and local neighborhoods in particular [50], but for our purposes we solely consider edge count between taxa as the defining classifier. As alluded to earlier, many trees do not provide distances along edges that would allow us to use total inter-species distance to classify other taxa as local or non-local neighbors. While these distances are crucial in many biological analyses [51], we can safely ignore them for our purposes as we are merely interested in a set of closest neighbors and not the actual distance between them. We begin by defining a maximum edge count

distance τ as the outer perimeter of a taxon’s local neighborhood. This distance can be adjusted to meet the particular study’s needs. In rooted trees, neighborhood sizes equal to half the average height of the tree have proven to be most useful in our tests on small biological datasets while results on larger datasets suggested that $\tau > 4$ rarely yield an improvement in distance score quality. It should be noted that defining smaller neighborhoods sacrifices a relatively small share of performance while selecting an excessively large neighborhood size very quickly reduces the quality of the resulting distance scores. It will become obvious that the smallest feasible size will be $\tau = 2$. Based on these observations, we can state that $\tau = 4$ is suitable for trees with at least 16 (2^4) taxa. For smaller trees, $\tau = 2$ is suggested.

Starting with the initial taxon we traverse up the tree for τ edges or until we arrive at the root. All visited nodes will be internal nodes and at each such node we fork a separate process that recursively traverses the node’s other child node and its respective children along as many as $\tau - \rho$ edges, where ρ is the edge path distance between the original taxon and the node at which the fork occurred. In a rooted tree, the root is treated equally to any other internal node, except that the upward walk obviously ends at the root. Since all taxa are necessarily leaves in gene trees, we need not traverse downwards from the original taxon. All taxon nodes that are visited along this recursive traversal are added to the list of local neighborhood taxa exactly once. This limitation is necessary to assure generalizability for phylogenetic networks and need not be implemented for regular phylogenetic tree structures as these will by their nature ensure that each leaf node will only be visited once. The size of the local neighborhood - and hence the number of taxa reported as local neighbors - depends on the actual tree structure. While the upper bound of the neighborhood size is defined by $2^{\tau-1}$, the neighborhood can actually

be empty. If a taxon’s local neighborhood within a tree is reported as an empty set, this tree/taxon combination must be excluded from any further calculations. Alternatively, a larger τ may be used to increase the local neighborhood’s size and thereby the probability of finding local neighbors. It should be noted that through this effect, local tree structure does affect any distances we may later calculate based on such a local neighborhood. While we are not interested in the overarching tree structure, local structural changes are relevant to our study and therefore this indirect measure of tree structure is useful and intended.

3.2.2 Species Displacement

In order to determine how much a given species’ placement differs between any two trees A and B, we can merely compare its local neighborhoods within those respective tree structures. We compare the two sets of neighboring taxa and score them based on how well they are conserved between the different samples. We begin by calculating the union of the two local neighborhoods and classifying each neighboring taxon based on the four potential outcomes:

- (a) Taxon is a local neighbor in both trees.
- (b) Taxon exists in both trees but is only a local neighbor in A.
- (c) Taxon exists in both trees but is only a local neighbor in B.
- (d) Taxon only exists in one of the two trees.

Taxa that do not exist in one of the two trees can obviously not be used to detect differences in neighborhoods. Taxa that are local neighbors in both trees indicate that the base species’ placement is rather similar in both trees while options b) and c) indicate dissimilarity in this placement. Depending on the particular research interests, different scoring matrices may be used, but in general b) and c)

will increase the distance score while a) should either be neutral or lower the distance score. The latter option may be problematic when comparing highly similar structures as distances could turn negative or become inaccurate if limited to positive values. In some applications it may be useful to assign different scores to b) and c) if the direction of change is of interest. In our larger study it is not. One might also consider assigning a distance value for d) in certain scenarios. The base species' displacement score is computed by summing up the different neighboring taxa's scores and normalizing for the size of the neighborhoods.

3.2.3 Species Displacement in Sets of Trees

We repeat the process of determining and analyzing a local neighborhood for the same species in all other trees of interest. While we can compute species specific distance scores between all possible tree pairings, most practical applications will focus on comparing a taxon's placements in a larger set of trees against a benchmark tree - most likely a consensus or species tree. In this context it should be noted that each of the specified neighborhoods may contain a varying number of taxa and the different neighborhoods may or may not share some or all taxa. Only pairs of trees with a sufficiently large intersection of local neighborhood taxa can be scored while some trees may not be scorable. It is apparent that any tree that does not have the base species in its namespace cannot be scored whatsoever. This process yields a list of individual-species distances (or a matrix if all trees are compared to each other) between the benchmark tree and all other scorable trees. Said list provides information about trees in which the analyzed species might be particularly displaced.

3.2.4 Tree Distance

Tree distance can be measured in a large variety of ways. However, most of them require edge distances and those that do not tend to overvalue small changes and undervalue incremental changes. For the aforementioned purposes we need a measure that detects small changes in small clades and is less susceptible to producing large distance scores based on structural changes elsewhere in the tree. We can use the described local neighborhood comparison approach to compute a score that meets these requirements. In the initial case we compare a pair of two trees A and B whose namespaces share some intersection V. We determine a local neighborhood for each taxon in V for both A and B. We then compare and score the two neighborhoods as described above. As we perform this procedure for all taxa in V, we obtain a list of displacement scores which can be aggregated in any fashion suitable to the purpose at hand. In our study a simple arithmetic average score proved useful. Since we are operating on the basis of local neighborhoods, we need not normalize scores for tree sizes. In this fashion we can derive a distance matrix across larger sets of trees. Some distances may not be scorable as some trees may have incompatible namespaces. However, based on the larger number of pairwise comparisons, a distance based tree estimation tool may be used to build a metatree showing the relationships, similarities and distances among the original set of trees.

3.2.5 Formal Algorithm Description

Let $T_1, T_2, \dots, T_n$ be the leaf nodes (taxa) in a phylogenetic tree. Let P_{T_x} be the parent node of T_x . Let S_{T_x} be a node's sibling. Let B_{T_x, T_y} be a bipartition (ancestral node) and nodes T_x and T_y be its children. Let N be an initially empty list of local neighborhood nodes. Let E_A be the list of all neighborhoods in tree A and let $E_{A,Q}$ be a list of taxon Q's local neighbors in tree A. Let D be the distance

between two trees.

FOR EACH tree F:

FOR EACH T_q as Q:

WHILE $\rho < \tau$

WITH S_q as R and ρ as r

REPEAT RECURSIVELY

with each child of R WHILE $r \leq \tau$

$r++$

IF R is a leaf node, add R to N

Set Q as P_Q

$\rho++$

Add N to E as E_{F_q}

FOR EACH taxon U that exists in both trees G and H

Set A as $E_{G,U}$ and B as $E_{H,U}$

FOR EACH ELEMENT Z in $A \cup B$

If Z not in A and B, increment D

3.2.6 Runtime for Larger Sets of Trees

When computing a metatree based on pairwise comparisons between trees from a larger set of trees, runtime is naturally quadratic in terms of tree count ϕ . Dynamic programming approaches are bound to fail since distances in tree space are not transitive [52] and therefore each pairing has to be analyzed individually. Runtime in terms of taxon count, however, remains linear.

Recall that $2^{\tau-1}$ is the largest possible neighborhood size defined by the number of taxa considered to be local neighbors. The overall runtime can then be quoted as $O(\phi^2 2^\tau \sigma)$. If we consider setting $\tau = \frac{h}{2}$ where h specifies the height

of the tree as suggested in the neighborhood analysis above, we would achieve an asymptotic complexity of $O(\phi^2 2^{\frac{\lg \sigma}{2}} \sigma)$ which approaches $O(\phi^2 \sigma \sqrt{\sigma})$. As we will see later, τ is bound to be very small and constant, approximating $O(\phi^2 \sigma)$. Given sufficient memory, runtime can significantly be reduced by computing the entire set of neighborhoods for all tree/species pairs once, identifying unique local neighborhood taxon sets and then computing distances based on those sets alone. The exact performance gain will heavily depend on the particular sets of compared trees. Its upper limit is the previously stated $O(\phi^2 - \sigma)$ while the lowest theoretical runtime approaches linear time at $O(1^2 \sigma)$ as ϕ approaches 1.

3.2.7 Runtime for Pairs of Trees

The presented algorithm for computing species displacement between two trees will use constant sizes of neighborhoods and hence perform a constant number of operations on any tree size given a constant τ . Runtime does increase exponentially with growing τ . However, τ is a constant parameter and is unlikely to be set to large values since large τ would negate the local, taxon-centric approach this algorithm is designed for. When comparing two trees based on the displacement of taxa in the intersection of their namespaces, runtime increases linearly with a growing intersection. In the worst case this means a runtime of $O(\sigma)$ where σ specifies the overall taxon count. Breaking with common convention, we use σ for tree size in terms of taxon count rather than the more common n to avoid confusion in later chapters when we use n in a different context. Local neighborhood analyses and comparisons are performed exactly once for each species and therefore any added species only adds $O(1)$. We can hence state that TDeft's runtime for comparing a pair of trees is linear in terms of taxon count.

3.2.8 Phylogenetic Networks

Significant existing research has proven the generalizability of tree distance methods such as the topological Robertson Foulds measure to generalized phylogenetic networks [53]. These proofs were necessary since the underlying metrics were developed specifically for trees and in many cases specifically for phylogenetic trees. The algorithm presented here is inherently generalizable as it does not require a tree to have any particular structure beyond the characteristics that define a phylogenetic tree and network. Each node must either be a labeled leaf specifying a particular taxon or an unlabeled bipartition. This holds true across phylogenetic networks [54] and therefore our algorithm can be analogously applied to generalized phylogenetic networks. It should be noted that the particular implementation used for our larger study and for the analyses presented herein, is limited to treelike structures. By allowing this limitation, we can achieve faster processing times without sacrificing any needed functionality.

3.3 SIMULATIONS

In order to test our algorithm, we have simulated a comprehensive dataset consisting of base trees of 8 different sizes containing 2^3 , 2^4 , 2^5 , 2^6 , 2^7 , 2^8 , 2^9 and 2^{10} taxa respectively. We construct each base tree by initializing a tree with a root and two leaf nodes and subsequently replacing a random leaf node in the tree with a bipartition. This bipartition receives the replaced leaf node as its left child while a new leaf node becomes its right child. We assign this new leaf node a new taxon and continue this process until all available taxa are assigned. For each such base tree we generate a set of 100 n-away trees using the popular subtree-prune-and-regraft method (SPR). We select a random bipartition, remove one of its children along with the entire subtree rooted in that child node. We then select a second node at random and replace one of its children with a new bipartition. This

bipartition receives the original bipartition as one child and the previously pruned subtree as its other child. In this fashion we obtain a 1-away tree. Subsequent iteration of the same procedure yield 2-away, 3-away and ultimately n -away trees. For our simulation we create 100 such n -away trees. This set of trees allows us to perform an n -away analysis by investigating how well TDeft, as well as several other metrics, can distinguish between a tree that is n steps removed from the base tree and one that is $n+k$ steps removed. When repeated for multiple independent simulations, this test yields a good measure for the quality of a measure's reported distance scores - especially when differentiating between incrementally different trees is important. Our dataset contains 1000 independent simulations of 100 n -away trees for each of the sizes. In total, our simulation created 800,000 trees. We used this dataset to answer two essential questions:

- What is the optimal neighborhood size for each tree size
- Which method(s) yield the best scores for each tree size

3.3.1 Optimal Neighborhood Size

The quality of TDEFT's distance scores is greatly influenced by the neighborhood size τ . This becomes evident when one considers the edge cases. For $\tau = 0$, all taxon neighborhoods will be empty. The same holds true for $\tau = 1$ since each taxon represents a leaf node in the phylogenetic tree and is hence only attached to one edge. This edge will lead to its parent which is - by nature of phylogenetic trees - guaranteed to be a bifurcation. As such, this node does not add any additional taxa to the empty taxon neighborhood which hence remains empty. The only theoretical exception to this condition would be an unrooted 2-taxon tree in which two taxa are connected by exactly one edge. In this hypothetical scenario, $\tau = 1$ would include the sole other taxon and hence all taxa in the tree. This

situation is irrelevant since a 2-taxon relationship can only be represented in exactly one way and therefore tree distance measures are inherently unnecessary for and unconcerned with relationships between fewer than 4 taxa. In all trees with 3 or more taxa, a taxon's parent is guaranteed to be a bifurcation and hence the smallest τ that yields a non-empty taxon neighborhood is $\tau = 2$ which may add a neighbor to the neighborhood if the taxon's parent's other child is a leaf node. If said child is, instead, another bifurcation, the neighborhood remains empty. As such $\tau = 2$ considers - at most - each taxon's nearest neighbor. This allows for a very fine grained analysis that will detect minute changes between trees at the leaf node level. It is, however, less apt at recognizing larger scale changes in which entire subtrees are relocated. As we increase τ , the resulting scores better represent these larger scale changes while sacrificing a certain amount of sensitivity to small local changes.

As we shall see later, very small τ are beneficial when working with very saturated problems in which relative incremental changes become very small, especially in smaller trees.

Increasing τ will eventually deteriorate score quality as too many taxa become part of the local taxon neighborhood. Let h be the height of the tree and consider a $\tau \geq 2h$, for example. In this scenario, every single taxon in the tree will be considered a local neighbor as every single edge in the tree will be traversed. As a result, every taxon's neighborhood encompasses the entire tree namespace and hence all neighborhoods are identical. This, obviously, makes any comparison impossible and produces distance scores of zero.

It is therefore evident that one must carefully select a proper τ to ensure valid and usable distance scores. TDeft allows the user to define the neighborhood size to serve any given study's needs, but it is expected that many researchers may not

have sufficient information to make a proper decision regarding τ in TDeft. We therefore need to establish a default value that is based on a generalized optimum. We will demonstrate this using a large scale n -away simulation.

The size of the optimal neighborhood primarily depends on two factors - tree size and n -distance. Consider, for example, a 4-taxon tree with taxa A,B,C,D. Such a tree has exactly three different possible confirmations as taxon A can be a sibling to either B, C or D and the remaining two taxa naturally become each other's siblings as well. With only four possible states, it is impossible to make four subsequent changes without arriving at a previously used structure. It should be obvious that an n -away analysis with $n > 3$ is hence impossible. Even at smaller n , distance classifications will be abysmal as there is no sufficient variation among possible trees. Larger trees naturally produce more potential confirmations and are hence better suited for n -away analyses. However, even in large trees, classification rates depend on n .

The farther removed a tree is from the original base tree against which it is compared, the more variations have already been exhausted and the fewer options remain. At large distances from the original tree, further changes are more likely to produce structures that are more similar to the original tree rather than more dissimilar. This is obviously misleading since it will produce smaller distance scores than the previous tree. As a result, the problem of $n+k$ -classification becomes harder with growing distance n . Ultimately the ability to classify tree separation correctly by reporting a larger distance score for a farther removed tree depends on the absolute tree size and the relationship between tree size and distance. For example, in an 8 taxon tree the results are already disappointing at $n=1$ and for the same tree at $n=10$ the deleterious effect of distance yields classification rates that approximate randomness.

Figure 6: n-away analysis 32 and 64 taxa at n=10

Larger trees are subject to the same underlying problem, but due to their size, the true change signature dominates the increasingly small amount of randomness. Ultimately it is the relationship between n-distance and the number of taxa that determines the overall possibility of correctly differentiating between n-away and n+k-away trees. As Figure 6 illustrates based on 32 and 64 taxon trees, classification rates in a 32 taxon tree exceed 90% at n=10 and in a 64 taxon tree they soon approach perfect classification at the same distance.

When we look at the same tree and increase the n-distance, the negative effect of increasing n-distance becomes evident. As the ratio between taxon count and n-distance decreases, the classification problem naturally becomes harder. Figure 6 (I) is an example of such harder problems and also illustrates that in such cases a neighborhood size of $\tau = 2$ tends to outperform larger neighborhood sizes. This holds true for all smaller trees as well as larger trees at very large n-distances.

To illustrate the optimal neighborhood size in larger trees, consider a 512 taxon tree at distance n=50. Under closer examination of the top 10% of classification rates in the area of convergence, one can detect that the optimal τ is now larger. In fact, $\tau = 4$ and $\tau = 5$ are both equally good candidates. Since the tree has

Figure 7: n-away analysis in a 1024 taxon tree at n=50

a height of $\log_2(512) = 9$, this is in line with our initial observation on a small biological dataset that implied that the optimal τ should be equal to half the height of the tree. However, when testing against larger simulated trees such as a 1024 taxon tree (Figure 7), it became evident that on larger trees, the optimum remains around $\tau = 4$. Since this value is also relatively robust to harder problems, we consider it a viable default value and base our further calculations on it.

3.3.2 Comparison against other metrics on identical namespaces

In our comparisons we build on the methodology established by Kuhner and Yamato to select the most promising tree distance metrics[26]. We also incorporate parts of Kuhner and Yamato’s comparison programs although we heavily adapted them to fit our study’s needs. We use our own, substantially more efficient computations for Robinson Foulds and, of course, for TDef. In order to allow for easy comparisons between the studies, we use the same nomenclature and refer to the various methods by the same name that Kuhner and Yamato used.

As previously alluded to, smaller trees only allow for proper separation between n-away trees and n+k-away trees for relatively small n and even with small n, results cannot be expected to be perfect. The different distance methods all still

yield reasonably accurate classification rates for a 10-away tree (Figure 8). In this scenario, Align yields the best results for most k , achieving near perfect accuracy for $k=40$. TDeft only yields the second-best results peaking at $k=36$ and 93% accuracy which declines with larger k . Align still performs best in a 64 taxon tree with $n=10$.

Align consistently performs slightly better for a 1024 taxon tree as well (Figure 9), although at that tree size ($\sigma = 1024$) its computational requirements become increasingly prohibitive as its processing time grows cubically. The same issue exists for the lower performing MAST and we hence had to exclude it from larger trees. We continued to compute Align in our analysis although it would likely not be a feasible alternative for real-world biological studies that are limited to more reasonable architectures than the system at our disposal for this comparison. For example, a single comparison between two trees took an average of 240 seconds for Align while Robinson Foulds could be computed in a mere three seconds. Node consumed 21 seconds while MAST required 470 seconds and TDeft was complete after 6 seconds. It should be noted that a single distance computation is not an ideal benchmark as there is significant overhead in the testing environment that will skew the results in favor of the slower methods. In a later chapter we therefore use sets of tree comparisons when analyzing runtimes. It is further possible that more efficient implementations of the various algorithms exist as we merely deployed those developed by Kuhner and Yamato. Due to MAST's and Node's inferior performance we consider any optimization of its implementation to be futile. It might be worth pursuing better implementations of Align.

Figure 8: 32-taxon 10-away tree analyzed using all 5 metrics

Figure 9: 1024-taxon 10-away tree analyzed using the 4 fastest metrics.

3.3.3 Thresholding

n-away analysis tests a method's ability to distinguish between trees that are n steps away from the base tree and trees that are k further steps removed from the same base tree. This ability to separate two similarly distant trees from each other is crucial in many phylogenetic applications, especially those that base further decisions on minute incremental changes. However, it is imperative that we define what minimal change between two trees we consider as sufficient to classify the two trees as being sufficiently different from each other. As Kuhner and Yamato have noted in their 2015 comparison of methods, many metrics are likely to yield identical distance scores for several different n -away trees. Yet other scores will report slightly different scores for the same set of trees - sometimes due to the distance algorithm, sometimes merely due to rounding issues. In this situation it is questionable whether we can really favor one method over the other since the differences between the two are minute. We therefore chose to introduce thresholding into our further analysis. We not only require that an $n+k$ tree have a higher distance score, but one that is higher by a minimum relative amount. When using a 10% threshold, for example, any $n+k$ -away tree with a score smaller than 110% of the n -away tree's score will be considered identical. This intuitively results in lower correct classification rates for each k as illustrated in Figures 10 through 11. Enforcing such a threshold also reduces some noise and has a smoothing effect on the classification graph. It is noteworthy that with large thresholds TDeft appears to perform better. This is an indication that its differentiation between n -away and $n+k$ -away trees is more pronounced and that it is hence more robust to larger distances.

Figure 10: Comparing metrics on a 64-taxon tree using a 0% threshold

Figure 11: Comparing metrics on a 64-taxon tree using a 10% threshold

Figure 12: Comparing metrics on a 64-taxon tree using a 20% threshold

Figure 13: Performance on a 64 taxon tree with 25% mismatched data

3.3.4 Comparison against other metrics on heterogeneous namespaces

A much more interesting - and more realistic - testing scenario involves heterogeneous namespaces. In the previous analysis we assumed that all trees either have identical namespaces or that larger, more heterogeneous trees have been pruned to arrive at such homogeneous subtrees on which the distance metrics are computed. This, however, is somewhat unrealistic since biological data are inherently heterogeneous and certain taxa are bound to be present in some, but absent from other trees. It is therefore important to analyze how different metrics perform on trees with a heterogeneous namespace. For this purpose we deploy the same dataset, but we randomly select 25% of all taxa in the base tree and replace them with other taxa that are not present elsewhere in the dataset. In this fashion we are able to test each metric's robustness to missing taxa and its ability to produce good distance scores in realistic comparison scenarios. We expected metrics such as the Robinson Foulds tree edit distance to perform worse in the presence of missing taxa and the resulting inconsistent namespaces as they are known to plateau rather early [29], making differentiation between n -away and $n+k$ -away trees more difficult. Purely structural methods should be more robust to inconsistent namespaces. To better measure the methods' response to these more challenging comparisons we used a threshold of 10% in our analysis. It should be noted that without this threshold,

the various metrics’s qualities are closer and more mixed than what can be seen with an imposed threshold. It might further be of interest that with substantially larger thresholds of $> 20\%$, TDeft’s advantage increases notably while the absolute performance of all metrics - including TDeft - deteriorates. At a distance of $n=10$ and a threshold of 10% , TDeft more quickly recovers the differences between n -away and $n+k$ -away trees than any other metric (Figure 13). At larger k , Align yields marginally better classification rates. However, at $k < 5$, Align’s correct classification rate is outperformed by random chance. Hence TDeft clearly emerges as the favorable method, especially considering its much lower computational requirements. As expected, Robinson Foulds - while by far the fastest method - yields unsatisfactory results in this test case. Figure 14 shows the missing data response test from chapter 2 repeated with TDeft as an added method for comparison.

3.3.5 Processing Times

As previously mentioned, computational requirements vary greatly between the different methods with some operating in logarithmic time while other have exponential requirements. Table 1 lists the average processing times we measured for the five methods studied herein. All times are quoted in seconds and reflect the time required to compute 100 independent comparisons. We use 100 comparisons to filter out noise introduced by other processes simultaneously operating on the same machine and to produce a more usable value for the fastest method that only requires very small fractions of a second for a single comparison. To illustrate the computational complexity, we report processing times for four trees - each four times as large as its predecessor. In this fashion we can clearly see TDeft’s approximately $O(\sigma)$ requirements at constant τ . Robinson Foulds can be computed faster than TDeft, while all other methods require more time. Align’s requirements grow particularly quickly with two 1024 taxon tree already taking more than a day

Figure 14: Response to Missing Data with a 10% Threshold

Table 2: Runtimes with $\tau = 4$

Method	16taxa	64taxa	256taxa	1024taxa
TDeft	0.72	3.48	14.48	55.23
Robinson Foulds	0.35	0.66	3.20	17.5
MAST	6.10	39.41	623	22,760
Node	1.62	7.81	116	2,152
Align	4.1	67	2,427	85,367

to compare. These processing times highlight the need to compromise between speed and accuracy in real world applications.

3.4 Results and Discussion

All distance methods have particular strengths and weaknesses, but they all face similar challenges when operating on biological data. When comparing larger sets of trees, the trees' taxon namespaces are likely going to vary greatly as these trees are often built using different loci from annotated sequence data. These data sources are inherently incomplete and therefore taxa that are present in one or any given number of trees may be absent from the remaining trees. We have shown that some methods - particularly Robinson Foulds - tend to score trees with different taxon namespaces as rather different, except when the same two trees are pruned to a set of trees on the intersection of taxa. This requires substantial pre-processing of heterogeneous trees before they can be compared against each other. Especially in larger sets of trees with very different taxon namespaces, this substantially adds to the computational requirements, likely eliminating the advantage that Robinson Foulds has in this area. Many other methods require similar pre-processing as they are unable to score trees on non-homogeneous sets of taxa. TDeft requires no such pre-processing and can accept and measure any two trees as long as their taxon namespaces intersect. Taxa which only exist in one of two compared trees are simply ignored and do not inflate the overall distance score.

Several distance measures rely on the distances between sets of taxa or other internodal distances in general - some do so exclusively while others deploy a combination of these distance based approaches and a variety of structural comparisons. All of these methods require branch lengths to be associated with all edges in the tree. As alluded to in an earlier chapter, consensus trees often do not provide branch lengths as a result of the employed tree reconciliation method. In the absence of branch lengths, many distance measures are unable

to compare these consensus trees - often species trees - against other trees. It is, however, especially these species trees that will be used as the base trees in many analyses. Hence there is a clear advantage to using versatile tree distance methods that do not require branch lengths. Several such versatile methods exist. The well-performing Align is one of them and TDeft also operates independently of branch lengths.

Incremental, often marginal changes are another major challenge for distance methods. It is often important to distinguish similar sets of trees from each other from a much farther distance. Differentiating between trees in such a distant forest requires a method to avoid early saturation as observed in many classic distance metrics [29]. Several of the purely topological methods have proven to achieve this rather well with Align consistently performing best. Of all the compared methods, Align remained the best option when a varying number of thresholds were imposed. Only TDeft was able to produce similarly good results with either method outperforming the other in different test cases.

Since tree comparisons rarely happen in a vacuum, real world applications will likely require large numbers of comparisons between larger sets of trees. Due to this fact, computational requirements play an important role in selecting a distance measure. Robinson Foulds has proven to consistently outperform all other methods in this regard while Align was consistently the slowest. TDeft is slower than Robinson Foulds but substantially faster than most other methods - notably Align which proved to be exponentially slower in our analysis.

Considering these different challenges and the different methods' performances we confidently suggest that TDeft presents a viable and in many cases superior alternative to the existing methods as it produces results that approach and often exceed those of the best method while requiring less computational power than

all but one of the tested methods.

3.4.1 Future Directions

In our comparisons we heavily rely on existing work by Kuhner and Yamato and we use certain parts of their code to compute the methods they compared in their study. We do so to maintain comparability between different studies, but we also acknowledge that there could be inefficiencies in the implementations as they were created with the intent of comparing scoring quality, not speed. It is, therefore, entirely possible that faster implementations for Align are feasible. Optimizing the other methods' implementations may change the outcome of our study on computational requirements by a constant factor, although these potential gains are likely dwarfed by the asymptotic gains achieved by the different algorithms themselves. The theoretical foundations of the various methods suggest that neither could perform in near linear time like TDeft does.

CHAPTER 4

GENE ASSOCIATION IN TREE SPACE (GATRS)

This manuscript is being prepared for submission to PLoS Computational Biology.

4.1 Abstract

Relatively few studies exist to establish associations between genes and morphological traits. These studies are mostly limited to single sites and to binary morphological characters and they are very susceptible to missing data. Here we present GATRS (Gene Association in TRee Space), a novel algorithm that can operate on continuous morphological observations such as fin size and is robust to missing data by operating in gene tree space, analyzing linkages between entire genes and morphological traits. We test GATRS on a comprehensive dataset of osteichthyes with 211 taxa and 2072 loci and produce 160 implied associations. In spite of the sparse amount of existing research, we are able to demonstrate the correctness of more than half of these associations, confirming the validity of our approach. We also report new discoveries for which no previous laboratory research exists and suggest that GATRS serve as a guide for future biological research studies.

4.2 Introduction

4.2.1 Motivation and Significance

Morphological characteristics are heavily influenced by environmental pressures [1], epigenetic control [2] and, of course, information encoded at the nucleotide level [3]. In this study we attempt to detect signals that can associate given genes with morphological characteristics in spite of the significant complexity of the overall system and the large amount of noise that is introduced by factors such as environmental pressures and epigenetic control. As many gene regions control early developmental patterns, they often affect multiple morphological traits [4]. Similarly, each trait may be influenced by larger sets of loci. The relationship between genes and morphologies is hence a complex n-m relationship in which any number n of genes influences any other number m of traits. Such complex

relationships cannot easily be modeled or explained. Determining which genes are responsible for particular morphologies can hence be very time consuming and expensive [5] as extensive laboratory experiments are necessary [6]. These experiments often involve animals, requiring even more resources and raising significant ethical concerns [7]. For example, in the widely used knockout study approach [8], researchers create mutant individuals by deactivating, deleting or mutating particular genes or loci in order to observe the effect on the organism [9]. This often produces animals with severe deformations or diseases that may be expected [10] or unexpected [11], which has led to a renewed discussion of the ethical tradeoff between animal welfare and research outcomes [12]. Other common studies aim to determine the effect certain external pressures have on organisms during their developmental stages. For example, certain fish populations in polluted waters display unusual and biomechanically disadvantageous [13] fin shapes that are not present in similar populations of the same species in cleaner waters [14]. This suggests that one or any combination of the compounds present in the polluted environment act epigenetically on fish development and produce disadvantaged adults [15]. To determine which gene expressions are affected – and to possibly link particular pollutants to their adverse effects – studies usually analyze gene expression in normal and abnormal individuals [16]. To achieve this, samples are treated with primary antibodies that specifically bind to the proteins coded by a particular gene. Secondary antibodies are used to visualize these primary antibodies using fluorescent properties [17]. As a result, the specimen can be visually analyzed to determine exactly where in a given tissue sample a certain gene is expressed. While very effective and commonly used, these studies are cost intensive as the primary antibodies must be created individually for each gene one wishes to visualize. Also, while a few genes can be visualized in the same sample, much

larger sets of samples would be needed for an uninformed search for expressed genes. Such studies can be much more efficient if one has a general idea of which genes should or might be expressed in a tissue sample and which expressions one might want to visualize. By computationally determining which genes appear relevant to given morphologies, one can better focus experimental analyses. This, in turn, has the potential to reduce the number of needed laboratory studies as much of the preliminary and intermediate work can be replaced by computational analyses and simulations [18]. Better experimental focus obviously saves time and cost and can thereby increase the pace of overall discovery as well. In this work we introduce and test a novel method to identify genes that are likely to have an influence on certain morphologies. Further studies can then be focused on these candidate genes.

4.2.2 Existing Approaches

Genome Wide Association Studies (GWAS) are the most commonly used approach to infer gene functionalities on a larger scale. GWAS operates on single nucleotide polymorphisms (SNP) or single site amino acid convergences and hence requires the exact loci to be available and properly mapped for all compared individuals [55]. Statistical approaches attempt to allow GWAS computation on data sets with missing data [56], but they may in certain cases lead to erroneous [57] and even positively misleading results [58], even when corrective measures are applied [59]. Traditionally, GWAS has mostly been applied at the population level, but computationally the approach should work analogously on data sets across different taxa. This opens up the opportunity to find cases of convergent evolution at the species level and could establish gene associations that cannot be seen at the population level since many traits might be too highly conserved between individuals of the same species. Obviously working across species will also add noise and,

in addition, many real world datasets will provide very sparse coverage as often only a small number of loci have been sequenced and annotated for each species. However, in spite of these challenges, certain GWAS methods should be considered as alternatives to the approach presented here. In recent years, many GWAS studies have attempted to link convergent evolutionary trends to positive selection on the molecular level. Such links could imply that the implicated locus directly or indirectly supports the development of the convergent traits. Some studies use a larger focus in an effort to find loci that may favor the development of traits that are beneficial in certain environments [60] while others aim to locate molecular sites that are responsible for specific traits [61]. One well discussed example focuses on the presence or absence of echolocation capabilities in mammals. These studies vary in their bioinformatics approaches and often yield inconsistent results. In their 2013 study, Parker et.al. report a significant convergence of certain genes for bats and bottlenose dolphins, both of which are echolocating mammals [62]. They particularly focus on genes that had previously been implicated in the development of hearing abilities in echolocating mammals. This result is somewhat surprising as one would expect the substantial differences between sonar systems in dolphins and bats [63] to be produced by more variable molecular bases. Dolphins' hearing system is also significantly adapted to subaquatic hearing [64] which again should be reflected in some molecular divergence. Indeed, a reanalysis of the aforementioned study found that no significant genome-wide protein sequence convergence for echolocation existed [65]. On the other hand, certain nucleotide convergences that have been shown between whales and bats [66] remain unchallenged. These apparent disagreements may be caused by insufficient morphological signals, the choice of data sets, excessive noise or the insufficiency of the deployed bioinformatics approaches, leaving unanswered the question whether molecular foundations for

morphological traits can accurately and reliably be determined [67]. A particular weakness of GWAS studies arises from the fact that they were mostly developed to detect binary conditions, such as the presence or absence of a particular trait. This is a direct result of the conceptual basis of correlating phenotypic convergence with convergence or at least convergent shift on the molecular level. This approach does not allow for the analysis of varying morphologies such as the relative sizes of limbs or the number of rays in a fish’s fins. For such cases, convergent evolution analysis – and hence single-site GWAS as a whole – is less useful. By focusing on particular sites, convergence studies are also very susceptible to missing data. While sequencing technology continues to improve at a rapid pace, missing data remain a common problem and hence any Bioinformatics approach should be robust to a certain amount of missing information. Our study does not operate on the molecular level but instead works in phylogenetic tree space. Since tree inferring methods are in themselves robust to missing data [68], this abstraction makes our method less susceptible to missing data while also considering the inferred distance between taxa rather than the convergence of particular sites. This is more likely to represent interaction between simultaneously changing sites and, as it is distance based, allows for the analysis of continuous (non-binary) traits. These benefits are achieved by detecting homologies across species that have developed similar traits to cope with similar external pressures. It should be noted that the resulting molecular convergences are typically considered disadvantageous to phylogenetic reconstruction as they have the potential to mask and override the underlying evolutionary signal [69]. This, however, only presents a problem when trying to determine the one true phylogeny based on either individual loci or whole genomes. We instead rely on sets of individual gene trees and a consensus tree among them. ASTRAL II - the consensus method used here [21] - has been shown to be robust

to missing, incomplete and partially inaccurate data [70], yielding a reasonably accurate phylogeny even in the presence of convergent adaptation [23]. It is in this context also noteworthy that our study does not aim to report any true phylogeny but rather operates on interarboreal discordances. Consequently this method will be robust to the misleading phylogenetic inference based on the non-phylogenetic signal [20] created by homologous adaptation [69] while using precisely that signal to link particular loci to the selected adaptations and traits.

4.3 Methodology

4.3.1 Data Sets

GATRS requires two distinct datasets to establish associations between loci and traits. First, it requires a set of gene trees or – more general - locus trees. These trees can be built on nucleotide sequences or protein sequences. Operating entirely in tree space, GATRS could also be used for trees built on entirely different characteristics present in different disciplines such as Astrophysics that have recently begun using phylogenetic tools for their purposes [71]. In this study we will focus on traditional phylogenetic analysis based on nucleotide sequences. GATRS also requires a second dataset describing morphometric observations. This dataset may theoretically include any number of traits or characters as long as they can be expressed in numerical values. These two datasets are distinct and independent, but they must contain an intersection of taxa as GATRS will compare a taxon’s displacements in one set versus the other.

4.3.2 Data Selection

In order to properly associate continuous changes in observed characters with changes in a locus, we needed a somewhat continuous dataset encompassing a large number of species that are based on a sufficiently similar body plans. Comparing species with diverging body plans would mask these effects and likely make them

untraceable as diverging body plans develop very different distinct characteristics [72]. Continuity within the data set is also crucial as incremental changes are most notable in very similar species. Mammalia, for example, would generally not be good subjects as extinction events have led to larger interspecies distances [73]. With fewer taxa remaining, neighboring taxa will likely display a larger number of morphological differences than neighbors in a more densely populated phylogenetic tree such as that of birds or fish [74]. Some insects would theoretically be reasonable choices although morphological observations tend to be difficult due to their size and the often extreme and nonlinear variability [75] between individuals produced by environmental pressures [76]. A particularly common dataset used in Bioinformatics is that of the drosophila family. This dataset is very well annotated and functions are known for a large number of genes. Working with fruit fly taxa also has the benefit of operating on a very small genome. Unfortunately there are only 12 taxa in this dataset which makes comparisons in a tree built on their nucleotide sequences very difficult and prone to random error. This problem is exacerbated by the fact that fruit flies are subject to large amounts of intraspecies variability that actually exceeds interspecies variability and hence makes many comparisons futile. It is obvious that we need a much larger dataset with distinct but close taxa. The most obvious choices of subjects for this study would come from one of two classes: either the increasingly well sequenced, annotated and studied birds (Aves) [77] or the most biodiversically rich class of vertebrates [78] - fresh water [79] or marine [80] fish (Pisces). For this study we chose to use an extensive osteichthyes dataset building on data from previous research.

4.3.3 Genetic Data

We performed our initial exploratory research using a published dataset on roughly 1,100 taxa and 21 loci [81] and then proceeded to obtain a much larger

dataset for our further analysis by querying GenBank for nucleotide sequences. We scanned approximately 5,000 loci of which we selected the 2,072 with at least 10% taxon coverage. The union of all taxon sets present within these trees yields roughly 1,400 taxa, of which we selected the 211 that are present in at least 10% of considered gene trees. In total, we are testing GATRS using approximately 100,000 distinct sequences.

4.3.4 Morphometric Data

To build a broad dataset containing consistent morphometric observations for a large number of taxa we queried FishBase [82] and assembled a collection of approximately 1,400 taxa. While obviously not all observations were available for all taxa, FishBase proved to provide good overall coverage for the both relative and absolute measurements. Relative measurements are reported as a function of total fish length which makes them easily comparable across taxa. In this category we collected the following values:

- Anal fin position
- Body depth
- Dorsal fin position
- Eye diameter
- Eye position
- Fork length
- Head length
- Pectoral fin position
- Pelvic fin position

- Anal fin length
- Caudal fin length
- Dorsal fin height
- Dorsal fin length

In addition to these easily comparable relative measures, FishBase also provides absolute numerical values for many observations. As some of these have very limited variability, they are not of much use for our further analysis and we chose to only use those marked with an asterisk (*). It should be noted that some traits would theoretically be excellent values to compare between taxa, but still had to be excluded from this study due to limited data availability.

- Anal fin count
- Anal fin soft rays *
- Anal fin spines *
- Aspect ratio of caudal fin *
- Dorsal fin count
- Dorsal fin soft rays *
- Dorsal fin spines *
- Dorsal finlets
- Gill rakers (for upper and lower limb separately)
- Lateral lines
- Pectoral fin soft rays *

- Pectoral fin spines *
- Pelvic fin soft rays *
- Pelvic fin spines *
- Pored lateral line scales
- Pre-anal vertebrae
- Total vertebrae *
- Scale rows (above and below lateral line reported separately)
- Scales around caudal fin
- Ventral finlets
- Scales on lateral line

4.3.5 Missing Data

As is the case in most biological datasets, our collection of nucleotide sequences is incomplete due to significant amounts of missing data. Since most available sequences are the result of particular research projects on certain groups of taxa and/or loci, virtually no locus is available for any large number of taxa and, inversely, not all loci can be expected to be available for any one taxon. GATRS is inherently robust to missing data as it operates in tree space and tree inferring methods themselves have proven to be quite robust to missing data [68]. Taxa missing from one gene tree can still be expected to be in their accurate location in the species tree as long as enough information is available in other genes [83]. However, we chose to reduce the overall processing time in our test scenario by removing loci with coverage of fewer than 10% of the taxa. We similarly removed taxa that were present in fewer than 10% of the locus trees.

4.3.6 Alignment

While there have been attempts to infer trees on unaligned sequences [84], the resulting trees are insufficiently accurate for our purposes [24]. We will therefore have to rely on algorithms that operate on aligned sequences. Alignment can be computationally expensive and the quality of the outputs varies greatly among the large number of existing sequence aligners [85]. Examples include MAFFT [86], Clustal Omega [87], Hidden Markov Model aligners [88], Kalign [89] and Mauve [90]. Alignment quality is typically scored by simulating nucleotide sequences with inserts and deletions. The simulation will yield the true alignment as well as the simulated sequences and an aligner's output can be compared against that ground truth. Comparative studies also yield the required computing time and generally one has to choose a method based on a compromise between speed and quality [91]. In building the gene forest for GATRS, we chose ClustalW, a widely used, highly efficient aligner that parallelizes well [87].

4.3.7 Gene Tree Inference

There are numerous competing gene tree inference methods such as PhyML [Guindon2005], RaxML [92], PAUP [93], IQ-Tree [94] and Mr. Bayes [95]. Their performances vary greatly and some are more suitable for certain applications than others [96]. Here we will use fasttree [97] as it is one of the least computationally demanding algorithms that yields acceptable results and that allows for parallelization to decrease required computing time.

4.3.8 Species Tree Inference

The assumed species tree is calculated by building a consensus tree based on all computed gene trees. Oftentimes one of the gene trees represents the consensus tree, but this is not necessarily the case. Approaches to consensus tree estimation

vary greatly, as do opinions about their validity and accuracy. In this research we used ASTRAL II, a coalescent based estimation method [21] that has consistently proven to be the most reliable and effective in our previous research and has also been proven to be one of the best methods in recent comparative studies [23]. Its particular strength arises from its robustness to missing data or misleading signals in individual gene trees [70]. Such misleading signals often arise as the result of insufficient data or an unfit evolutionary model during the gene tree estimation phase. It is important to keep in mind that the true phylogeny is less important for the further analysis and hence such misleading signals at the gene tree level are less problematic. It is also worth to remember that this assumed species tree should not be accepted as a true species tree for other purposes as it is built on a subset of loci that may or may not reflect actual phylogenies.

4.4 Gene Association in Tree Space (GATRS core)

4.4.1 Morphometric Space

GATRS establishes associations between loci and morphometric characters by comparing taxa’s placement in various trees with the extent to which they display a particular trait. This comparison between genetic tree space and morphometric observations requires displacement scores to be computed for each taxon in both realms. The process is rather obvious in morphometric space. For each given taxon, we establish an overall displacement score or oddity score. This is achieved by computing how different said taxon’s expression of a certain trait is from the average expressed by all taxa for the same trait. This displacement is then computed for all available traits and the average of all of these scores yield information how far removed the taxon is from a hypothetical average taxon that scores exactly average on all trait expressions. This score in itself yields no usable information, but it is useful in determining which characteristics of a given taxon are partic-

ularly pronounced as we perform a second pass over the computed displacement scores for the various traits and normalize them by the average oddity score we just derived. This produces a displacement score for each taxon within each trait that holds information as to how unusual a taxon’s expression of a certain trait is. The larger this score, the more said taxon stands out among its peers when one observes a given characteristic.

4.4.2 Gene Space

The approach in tree space is essentially very similar, although the methods to derive displacement scores are different due to the nature and structure of the data in tree space. We need to calculate a score that holds information as to how unexpected a taxon’s placement is in a given gene tree. While in morphometric space we can simply make use of average observations across all taxa for the same trait, we need a different benchmark in tree space. The tree space equivalent to such an average value is a consensus tree that represents the most parsimonious hypothetical tree that best agrees with as many gene trees as possible. As previously alluded to, we derive such a consensus tree by first building the individual gene trees using *fasttree* [97] and then creating a consensus tree using *ASTRAL II* [21].

Once we have inferred all gene trees and the corresponding consensus tree, we can compare a taxon’s displacement within any given gene tree against its placement in the consensus tree to determine how unexpected its placement in the gene tree is. We do so by using *TDEFT*, a newly developed tree distance measurement method that not only reports tree distance scores, but also provides displacement scores for each taxon in the tree [98]. *TDEFT* has been shown to be superior to most of the existing tree distance methods compared by Kuhner and Yamato [26]. Their study particularly investigated performance differences between topological

methods and those that partially or fully rely on branch lengths. However, since we are comparing gene trees against consensus trees, branch lengths are not available for our purposes which limits us to the purely topological methods. The majority of these – like Trip [48], MAST [99] and Align [27] are mostly concerned with differences in the overall structure of entire rooted trees. Some, like the Robinson Foulds algorithm [100] tend to plateau too soon and can hence not reasonably identify incremental changes [29]. Ultimately Align has emerged as the best distance measure in both Kuhner and Yamato’s study as well as our own exploratory research. However, Align’s computational requirements make it a far inferior choice for applications such as GATRS that rely on large numbers of comparisons between large trees. Furthermore, we do require displacement scores for each taxon to perform the displacement correlation within GATRS and TDEFT is the only measure that can provide such scores. Once we compute the displacement scores of all taxa for all available genes, we normalize these values by the average displacement of all taxa for each given gene. This process reduces noise introduced by more entropic trees that are far removed from the consensus tree.

4.4.3 Comparison between Morphometric and Gene Spaces

After completing these two separate computations in phylogenetic tree space and in the realm of morphometric observations, we can assemble pairs of lists of displacement scores. Any one such pair will list the displacement scores for a list of taxa in a given gene tree on one side and the same taxa’s displacement scores for a particular trait on the other. This yields two aligned data streams that can be analyzed for correlation. As GATRS does not aim at identifying particular associations, we do not need to impose a threshold, but can merely report correlations and p-values for each genes/trait pair. Naturally all p-values need to be corrected for multiple tests and here we chose to do so using the Benjamini-

Hochberg method [101] . It should be obvious that pairings with high p-values are of no particular interest, except when testing against a true or simulated set of observations that are known to be insignificant or random. In order to establish directionality and to reduce the likelihood of false positives due to gene tree entropy, we further perform a clustering analysis. Using TDeft's neighborhood structure we select neighbors for each taxon in a gene tree and compare the observations for all examined traits among the different neighbors. Specifically, we compute the difference in observed trait values for the reference taxon and each of its neighbors and normalize the obtained distance by the entire range of values observed for the same trait across all taxa in the same gene tree. This will yield a trait distance score between 0 (identical) and 1 (most different) for each neighbor. By adding the trait distance scores for all neighbors, we arrive at a clustering score between 0 (perfectly clustered) and an upper bound of $2^{\tau-1}$ if all neighbors are as different as theoretically possible. This upper bound is determined by the largest possible neighborhood size for any given τ as introduced in an earlier chapter.

Let N be the number of taxa in a tree

Let V_n be the set of taxon n 's neighbor taxa

Let T be the set of observations for a given trait

Let T_x be the trait value observed for taxon x

$$c = \frac{1}{N} \sum_{n=0}^N \sum_{v=0}^{V_n} \frac{abs(T_v - T_n)}{max(T) - min(T)}$$

By computing the mean of every taxon's neighborhood's clustering value, we obtain a cluster score for the entire tree. A perfectly clustered tree might produce the lowest theoretically possible score of 0, while trees in which taxa with very similar observed values don't cluster well will produce higher score values, up to $2^{\tau-1}$. Figure 15 illustrates the distribution of c-values among statistically significant associations.

Figure 15: Distribution of c-values among statistically significant associations

It is important to note that discord between gene trees and species trees can exist for a wide variety of reasons. Examples include evolutionary processes such as Incomplete Lineage Sorting (ILS), Horizontal Gene Transfer (HGT) [102] and Horizontal Transposon Transfer (HTT) [103]. ILS most often arises when speciation occurs during an incomplete mutation. While a population undergoes such a mutation, any descendent species might inherit both the pre-mutant and the mutant alleles and then later pass either one of those on during future speciation events. This will obviously mislead phylogenetic analyses built on any such affected locus. It may, however, provide insight into that locus' relation to the original mutation. HGT, on the other hand, describes a process in which genetic material is transferred between contemporary organisms, providing equally misleading phylogenetic clues. This is relatively rare in eukaryotes, but a common occurrence in bacteria [104]. Eukaryotes are, however, likely to experience a similarly problematic process called HTT that refers to the transfer of nucleotide sequences within the same organisms, but between different genes [105]. This could be particularly challenging for phylogenetic analyses as it might lead to erroneous gene annotation and consequently

nonsensical phylogenetic inferences. All of these factors are likely to produce gene tree discord – some of which may hold important non-phylogenetic clues, some simply adds noise. Gene trees might also differ from each other for reasons not directly rooted in the quality or nature of the underlying data. Trees are built using a variety of methods and virtually all of them rely on some type of evolutionary model. Depending on the sophistication of the underlying model as well as its fit for the particular evolutionary realities that have driven the true evolution that has produced the analyzed data, these methods may yield more or less accurate estimates of evolutionary history [20]. Additionally, the mere fact that the species tree is inferred on the basis of non-homogeneous gene trees automatically produces discord unless all genes produce the same tree which is highly unlikely. Factors such as these introduce a large amount of random noise into the observed tree space. However, maintaining the underlying evolutionary signal in the presence of such noise is a well-documented strength of the consensus tree method [106] as opposed to building species trees on gene read concatenations [107]. Hence this signal should remain and be detectable. In fact, since we compare displacements of two trees within the same tree space with respect to the reconciliatory consensus tree, the introduced noise should affect both comparisons equally and thereby be largely cancelled out [108]. There is, in fact, evidence that in larger tree spaces, discord of individual trees caused by ILS or HGT might improve the accuracy of comparative analyses [70]. This initially counterintuitive result concurs with observations of other computational estimations of evolutionary processes such as gene tree inference based on incomplete data sets. In these situations, missing data has also shown to not affect or even improve the overall accuracy of estimations produced by coalescent models [83]. Hence GATRS is able to produce a reasonably strong signal.

4.5 Results

We successfully verified GATRS' validity and performance by consulting with a group of functional morphologists and marine biologists to evaluate associations implied by GATRS and to compare them to existing literature and research. It is important to note that there was no training data and that no ground truth or known associations were used as inputs. The GATRS core only accepts gene trees in newick format as well as a list of observed values in CSV format. The former obviously only contains taxon names and their position within a gene tree, each record in the latter provides a taxon name, an observation name and an observation value. GATRS operates without prior knowledge of any kind and bases its associations purely on the supplied unannotated data.

After running GATRS on the previously described osteichthyes dataset, we corrected all significances for multiple tests using a Benjamini Hochberg correction allowing for a false discovery rate (FDR) of 1% which resulted in a significance boundary of 0.0041 [101]. We then grouped the various available traits into closely related groups that describe the caudal fin, the dorsal fin, the anal fin and properties of head and eye. Here we introduce the highest ranking gene associations for each group and document any known interactions that may exist in literature. As functional annotation of genes is sparse and complicated by the pleiotropic nature of the existing pathways, one would expect a large number of currently unverifiable computed associations. However, the same pleiotropic situation and sparse annotation also make it much more unlikely that we would be able to find support for any of our suggested gene associations unless that relationship is quite strong. This, in turn, exponentially increases the weight of each verifiable association. In the tables in this chapter we will denote gene associations that could clearly be proven using existing literature with a double asterisk (**) and such associations

for which we could determine a likely pathway with a single asterisk (*).

4.5.1 Caudal Fin

Adaptive evolution of caudal fins can be expected as a response to environmental pressures as caudal fin shape and flexibility directly influence swimming speed [109] which is important to secure food and avoid predators. We found two genes associated with caudal fin length: *smad2* ("Mothers Against Decapentaplegic") and *eif3g* ("Eukaryotic Translation Inhibition Factor 3 Subunit G") with p-values of 0.00004 and 0.001 respectively. *smad2* is mostly studied in the context of immunology [110], but there is evidence that the *smad* family acts as effectors of the bone morphogenesis protein (*bmp*) signaling pathway thus influencing both fin development and regeneration [111]. Zebrafish mutants for *smad2* also show abnormal caudal fin phenotypes, further supporting the influence of this gene [112]. There appears to be no known pathway linking *eif3g* to caudal fin size at this time. The only significant association reported for caudal fin aspect ratio involved gene *tfec* ("Transcription Factor EC") with a p-value of 0.0003. Expression of *tfec* in various fins is well documented in zebrafish, as is its expression on both dorsal and ventral regions of the posterior region of the embryo and caudal peduncle [113] during the critical time of fin fold reabsorption and caudal fin formation.

Table 3: Caudal Fin Associations

Locus	Trait	p-value	c-value
<i>smad2</i> **	Caudal Fin Length	0.00002	0.0296
<i>eif3g</i>	Caudal Fin Length	0.0001	0.0301
<i>tfec</i> **	Caudal Fin Aspect Ratio	0.00003	0.0368

4.5.2 Anal Fin

A total of 33 associations were made between loci and anal fin position. Interestingly, *nuf2* ("Kinetochores Protein NUF2") has been a locus of interest in

cancer suppression research [114], but there appears to be no study confirming or rejecting its effects on fin position. *slc7a6os* ("Solute Carrier Family 7 Member 6") has been shown to affect somite morphology and – particularly ventral – body shape [115] which would in turn likely change anal fin position, although no exact pathway is known. Knockout studies have shown that *arghef19* ("Rho Guanine Nucleotide Exchange Factor 19") affects neural tube closure [116], but no direct pathway for fin position could be established. *acd* ("Adrenocortical Dysplasia") has been confirmed to affect ventral fin folds and hence positions of anal fins [117]. *cpsf3* ("Cleavage and Polyadenylation Specific Factor 3") appears to influence the position of multiple fins so an association with the anal fin is likely [118]. There is evidence that *rad52* ("DNA Repair Protein 52") alters overall body curvature which could certainly affect the anal fin [119].

Table 4: Anal Fin Associations

Locus	Trait	p-value	c-value
<i>mpnd</i>	Pre-anal length	0.00009	0.046
<i>nuf2</i>	Pre-anal length	0.00001	0.0552
<i>rad52</i> *	Pre-anal length	0.00025	0.056
<i>cpsf3</i> *	Pre-anal length	0.00021	0.059
<i>manba</i>	Pre-anal length	0.00022	0.059
<i>arhgef19</i> *	Pre-anal length	0.00006	0.0609
<i>slc7a6os</i> *	Pre-anal length	0.00001	0.0616
<i>acd</i> **	Pre-anal length	0.00007	0.0686
<i>prmt5</i>	Pre-anal length	0.00009	0.083

4.5.3 Dorsal Fin

pth2 ("Parathyroid Hormone 2") is involved in calcium metabolism and knockout experiments in zebrafish have produced an upwards curled trunk that would significantly affect dorsal fin formation [120]. Changes in *copa* ("Coatomer Protein Complex Subunit Alpha") have been shown to disrupt normal notochord development and segmentation [121]. Differences in segmentation obviously affect dorsal

fin formation. *sel1l* ("Suppressor of *lin12*") has known implications for fin formation and knockdown mutants show altered dorsal fin folds, confirming this locus' importance for dorsal fin position [122]. *slt* ("Soluble Lytic Murein Transglycosylase") affects curvature of the posterior section and thereby likely influences dorsal fin position as well [123]. *akt2* ("Threonine Protein Kinase 2") has been shown to affect dorsal fin formation and morphogenesis [124]. The dorsal fin lepidotrichium is known to be controlled partly by *bmp1* ("Bone Morphogenic Protein 1"), making that locus' association with dorsal fin rays unsurprising [125]. While *acd* has only been proven to affect ventral fin folds, there is no evidence that it does not also influence the position of the dorsal fin as these tend to be controlled by the same genes [117]. *carf* ("Calcium Responsive Transcription Factor") affects fin formation and regeneration and mutations cause structural changes in the dorsal fin region, making it a likely contributor to dorsal fin position [126].

Table 5: Dorsal Fin Associations

Locus	Trait	p-value	c-value
<i>tfec</i>	Dorsal fin soft rays	0.00008	0.0368
<i>col</i>	Dorsal fin soft rays	0.00001	0.0401
<i>bmp1</i> **	Dorsal fin soft rays	0.0005	0.048
<i>acd</i> *	Dorsal fin soft rays	0.0007	0.054
<i>akt2</i> **	Pre-dorsal length	0.00011	0.059
<i>ndp</i>	Dorsal fin soft rays	0.00011	0.062
<i>smg6</i>	Dorsal fin soft rays	0.00001	0.0621
<i>pth2</i> *	Pre-dorsal length	0.00002	0.0659
<i>ddc</i>	Dorsal fin soft rays	0.00001	0.0672
<i>snx7</i>	Dorsal fin soft rays	0.00013	0.068
<i>cdkn2aip</i> (<i>carf</i>) *	Pre-dorsal length	0.0007	0.071
<i>copa</i> *	Pre-dorsal length	0.00004	0.0721
<i>sel1l</i> **	Pre-dorsal length	0.00006	0.0787
<i>cos</i>	Pre-dorsal length	0.00005	0.0789
<i>ctnnb1</i>	Pre-dorsal length	0.00011	0.083
<i>slt</i> *	Pre-dorsal length	0.00010	0.086

4.5.4 Pelvic and Pectoral Fins

Among the roughly 2000 examined loci, no sufficiently significant association was found for pelvic or pectoral fin position or size. It is likely that these fins are influenced by too many loci within and beyond our study to create a sufficiently strong signal. We should also note that neither of these two observations displayed a large diversity. As the observed values become very close to each other, differences are harder to detect and likely overpowered by noise.

4.5.5 Pre-Orbital Length

The position of a fish's eye can drastically influence its competitiveness in certain habitats. It is therefore likely to converge among taxa that generally display similar behaviors and exist in similar environments. Pre-orbital length also presents a good proxy for overall head shape. *tmem165* ("Transmembrane Protein 165") has been shown to affect cartilage formation and the general shape of a fish's head as well as the eye's position within it [127]. *tfap2a* ("Transcription Factor for Advanced Enhancer Binding Protein 2 Alpha") has been shown to affect cartilage formation in the head [128], skull [129] and jaw formation [130] as well as eye size [131], all of which are directly related to pre-orbital length. *cpsf3* and *rpl9* ("Ribosomal Protein L9") phenotypes also display pathological head shapes and eye sizes [118]. *agbl5* ("ATP Binding Protein Like 5") has been shown to affect at least the eye's size and to have an effect on overall body shape [132].

4.6 Discussion

Traditionally, associations between loci and morphological traits have been mostly limited to binary conditions such as the presence or absence of certain features. This restriction has allowed studies to identify particular single site polymorphisms that tend to converge to a particular amino acid or at least shift towards

Table 6: Associations With Eye Position

Locus	Trait	p-value	c-value
tmem165 **	Pre-orbital length	0.00001	0.027
sf3a3	Pre-orbital length	0.00005	0.031
agbl5 *	Pre-orbital length	0.00005	0.034
pofut1	Pre-orbital length	0.00009	0.035
rpl9 **	Pre-orbital length	0.00003	0.037
ppih	Pre-orbital length	0.00009	0.037
tmem165	Pre-orbital length	0.00056	0.038
tfap2a **	Pre-orbital length	0.00001	0.039
cpsf3 **	Pre-orbital length	0.00003	0.039

such amino acids with similar properties. There are, however, many traits that cannot easily be expressed in a binary fashion. While it is easy to classify whether a taxon has echolocation abilities or not, one needs a continuous variable to describe dimensions such as fin length or position. Traditional studies are not able to accommodate such continuous trait descriptors while GATRS was specifically designed to take advantage of these non-binary traits and to associate them with candidate genes that are likely influencing such given traits. Generally, existing studies also focus on single site mutations and might hence miss co-mutations of different sites within a gene. The interactions between different simultaneously changing sites might either amplify or squelch each other’s effect, making it likely for analyses to miss true associations that were masked or to over-report those that were amplified. Genes – and more generally loci – influence protein folding and thereby function as a whole and it is imperative to consider them holistically when trying to find loci that have an influence on a particular trait. Once that locus has been implicated, more narrow studies can be used to analyze the underlying mechanics. GATRS serves as a guide for these more specialized analyses, suggesting loci that may be of interest for further research. In doing so it reduces the need for costly laboratory studies to narrow down the focus of functional anal-

yses. GATRS also addresses the problem of missing data which presents a major challenge for traditional studies. Missing data can be of particular concern when analyzing specific sites, but tree inferring methods have been proven to reconstruct accurate relationships in spite of incomplete data. This is a direct result of the aforementioned co-mutation of various sites within a locus. By operating in tree space, GATRS inherits this robustness to missing data and also requires much smaller amounts of input data than any study working on nucleotide sequences would.

We have tested GATRS on a comprehensive dataset including more than 1,000 osteichthyes (1,400 taxa including outgroups) and more than 5,000 loci. After eliminating loci with low taxon coverage ($< 10\%$), we included 211 taxa and 2,072 loci in our analysis. GATRS produced 160 assumed gene associations which we grouped into distinct parts of the fish body before analyzing the most significant associations for each group. Of the 36 considered associations, our literature review was able to establish proven relationships for ten and likely pathways for an additional ten. Those associations that could not be connected to existing literature are comprised of false positives and, more interestingly, true positives for which no previous research exists. There was no evidence in the reviewed literature that suggested any false positives among the reported associations.

Since GATRS itself does not use any biologically informed substitution models, it can easily be ported into other domains to solve a wide range of problems that can be modeled as phylogenetic trees. One such area in which phylogenetics are gaining popularity is the field of astrophysics [71]. It would be quite interesting to explore applications for GATRS in such other domains, but our method might also open up new possibilities in more traditional phylogenetics - perhaps most importantly in epidemiology and in various tumor studies.

The fact that we were able to verify more than half of the investigated associations using the inherently sparse existing research findings on the subject clearly proves that GATRS yields valid and significant results that can be used to guide further investigations into the suggested associations between loci and morphological traits.

CHAPTER 5

DISCUSSION

Associating genes with morphological traits is a relatively new field and most existing approaches have tried to connect binary characters such as the presence or absence of certain anomalies or features to local convergences in nucleotide sequences, often focusing on single site polymorphisms [57]. While these approaches are promising, they are very limited in their applicability as they do not allow for continuous observational characters such as the size of a particular body part to be linked to genetic changes. In the same fashion these studies are also not strong enough to link multiple simultaneous changes within a given locus to a particular trait as they do not consider individual mutations in context [61]. The interactions between different simultaneously changing sites might either amplify or squelch each other's effect, making it likely for analyses to miss true associations that were masked or to over-report those that were amplified. Genes – and more generally loci – influence protein folding and thereby function as a whole and it is imperative to consider them holistically when trying to find loci that have an influence on a particular trait [133]. Once that locus has been implicated, more narrow studies can be used to analyze the underlying mechanics.

GATRS, the new association method introduced in this dissertation, takes entire loci into account holistically and links their relative disagreement with the underlying phylogeny to relative changes in observational dimensions. It does so by operating entirely in tree space, which leverages two primary computational advantages - it is relatively robust to missing data and it requires substantially smaller amounts of input data than any method operating directly on the nucleotide level would.

Computing the relative discord values between gene and consensus trees requires a tree distance calculation method that can quickly process large numbers of comparisons and yield accurate differentiations between relatively similar trees gene trees that themselves are potentially far removed from the species tree.

Selecting the appropriate tree distance method has important implications for the quality and the feasibility of any study. Several distance measures rely on the distances between sets of taxa or other internodal distances in general - some do so exclusively while others deploy a combination of these distance based approaches and a variety of structural comparisons [26]. All of these methods require branch lengths to be associated with all edges in the tree. As alluded to earlier, consensus trees often do not provide branch lengths as a result of the employed tree reconciliation method [21]. In the absence of branch lengths, many distance measures are unable to compare these consensus trees - often species trees - against other trees. It is, however, especially these species trees that are used as the base trees in many analyses. Hence there is a clear advantage to using versatile tree distance methods that do not require branch lengths. Several such versatile methods exist and there is existing research that provides a solid comparison of popular distance calculation methods' performance in an incremental change analysis [26].

Incremental, often marginal changes are a major challenge for distance methods. It is often important to distinguish similar sets of trees from each other from a much farther distance. Differentiating between trees in such a distant forest requires a method to avoid early saturation as observed in many classic distance metrics [29]. Several of the purely topological methods have proven to achieve this rather well with Align consistently performing best. The comparison study presented herein built on exiting research[26] and added score thresholding to test their robustness to noise. It also comprehensively tested each method's performance when ana-

lyzing trees with limited taxon intersections. When comparing large sets of gene trees, taxon namespaces are rarely consistent and hence this ability is crucial for GATRS. The study's results show that there is no one generally superior method, but that one has to carefully select a method based on the nature of the analyzed data as well as the available computational resources. The latter aspect is rather crucial in this study as GATRS performs large numbers of comparisons and slower methods such as Align would be prohibitively expensive in computational terms although they generally yield the best distance scores. Align also proved to be more susceptible to incomplete data or heterogeneous taxon namespaces. In the absence of large taxon intersections between trees, Align lost its superiority and yielded increasingly disappointing results. It still outperformed most other metrics, but considering its much higher computational requirements, this relative gain in distance score quality makes it a less desirable choice for GATRS. Interestingly, MAST proved to be exceptionally robust to this test. While it was typically outperformed by most other methods on near-complete sets of data, it proved to be clearly superior when operating on small taxon intersections. MAST, however, was still too slow for GATRS. Robinson Foulds' computational requirements are much more amenable to GATRS' requirements, but it only yielded excellent scores for near-complete datasets. In the presence of incomplete taxon namespaces across trees, Robinson Foulds was easily outperformed by all other measures and quickly dropped below a 50% correct classification rate. Consequently it was not a good choice for GATRS either. Ultimately it became clear that none of the existing measures truly served the particular purposes of this new gene association method and hence an entirely new method had to be developed.

This study introduced TDeft, a novel tree distance algorithm that computes distance scores based on the relative displacements of individual taxa within one tree

versus another. TDeft has been shown to produce distance scores of a quality that often approximates that of the best existing methods while only requiring slightly more computational resources than the fastest – and inferior – existing method. When the previously described comparison study was repeated with TDeft added as an additional competitor, it was confirmed that TDeft provides an unparalleled balance between speed and quality of scores when operating on either complete datasets or in situations of considerable amounts of missing data. This makes TDeft the method of choice for GATRS.

GATRS was tested on a comprehensive dataset including 211 taxa and 2,072 loci. GATRS produced 160 assumed gene associations, the 36 most significant of which were compared to existing literature. It is important to note that GATRS does not use any prior knowledge or any sort of training data but only accepts gene trees and morphological observations as separate input datasets. It then derives its gene associations purely on those unannotated sets of data without any knowledge of the potential ground truth. It requires no training whatsoever.

As GATRS is a novel approach using biological data, one would expect to only find a small number of verified associations in existing research. GATRS should be expected to yield a small number of linkages that can be supported using exiting studies while a considerable number of presumed associations may need to be studied in a laboratory setting. Within this study, ten associations (27.8%) could be credibly verified using exiting literature and likely pathways could be established for another ten (27.8%). There was no evidence in the reviewed literature that suggested any false positives among the reported supposed associations. The fact that more than half (55.6%) of the investigated associations could be verified using the inherently sparse existing research findings on the subject and none could be disproven clearly proves that GATRS yields valid and significant results that can

be used to guide further investigations into the suggested associations between loci and morphological traits for which no current research exists. In doing so, it reduces the need for costly laboratory studies to narrow down the focus of functional analyses.

LIST OF REFERENCES

- [1] J. A. Granek and P. M. Magwene, “Environmental and genetic determinants of colony morphology in yeast,” *PLoS Genetics*, vol. 6, no. 1, 2010.
- [2] F. Balao, O. Paun, and C. Alonso, “Uncovering the contribution of epigenetics to plant phenotypic variation in Mediterranean ecosystems,” pp. 38–49, 2018.
- [3] F. Bertolini, G. Schiavo, S. Tinarelli, L. Santoro, V. J. Utzeri, S. Dall’Olio, L. Nanni Costa, M. Gallo, and L. Fontanesi, “Exploiting phenotype diversity in a local animal genetic resource: Identification of a single nucleotide polymorphism associated with the tail shape phenotype in the autochthonous Casertana pig breed,” *Livestock Science*, vol. 216, pp. 148–152, 2018.
- [4] H. Romdhani, H. Hwang, G. Paradis, M.-H. Roy-Gagnon, and A. Labbe, “Pathway-Based Association Study of Multiple Candidate Genes and Multiple Traits Using Structural Equation Models,” *Genetic Epidemiology*, vol. 39, no. 2, pp. 101–113, 2015. [Online]. Available: <http://doi.wiley.com/10.1002/gepi.21872>{%}5Cn<http://www.ncbi.nlm.nih.gov/pubmed/25558046>
- [5] R. M. Piro and F. Di Cunto, “Computational approaches to disease-gene prediction: Rationale, classification and successes,” pp. 678–696, 2012.
- [6] A. Bradley, K. Anastassiadis, A. Ayadi, J. F. Battey, C. Bell, M. C. Birling, J. Bottomley, S. D. Brown, A. Bürger, C. J. Bult, W. Bushell, F. S. Collins, C. Desaintes, B. Doe, E. Aris, J. T. Eppig, R. H. Finnell, C. Fletcher, M. Fray, D. Friendewey, R. H. Friedel, F. G. Grosveld, J. Hansen, Y. Hérault, G. Hicks, A. Hörlein, R. Houghton, M. Hrabé De Angelis, D. Huylebroeck, V. Iyer, P. J. De Jong, J. A. Kadin, C. Kaloff, K. Kennedy, M. Koutsourakis, K. C. Kent Lloyd, S. Marschall, J. Mason, C. McKerlie, M. P. McLeod, H. Von Melchner, M. Moore, A. O. Mujica, A. Nagy, M. Nefedov, L. M. Nutter, G. Pavlovic, J. L. Peterson, J. Pollock, R. Ramirez-Solis, D. E. Rancourt, M. Raspa, J. E. Remacle, M. Ringwald, B. Rosen, N. Rosenthal, J. Rossant, P. Ruiz Noppinger, E. Ryder, J. Z. Schick, F. Schnütgen, P. Schofield, C. Seisenberger, M. Selloum, E. M. Simpson, W. C. Skarnes, D. Smedley, W. L. Stanford, A. Francis Stewart, K. Stone, K. Swan, H. Tadepally, L. Teboul, G. P. Tocchini-Valentini, D. Valenzuela, A. P. West, K. I. Yamamura, Y. Yoshinaga, and W. Wurst, “The mammalian gene function resource: The International Knockout Mouse Consortium,” *Mammalian Genome*, vol. 23, no. 9-10, pp. 580–586, 2012.

- [7] D. G. Porter, “Ethical scores for animal experiments.” *Nature*, vol. 356, no. 6365, pp. 101–102, 1992.
- [8] J. Z. Tsien, D. F. Chen, D. Gerber, C. Tom, E. H. Mercer, D. J. Anderson, M. Mayford, E. R. Kandel, and S. Tonegawa, “Subregion- and cell type-restricted gene knockout in mouse brain,” *Cell*, vol. 87, no. 7, pp. 1317–1326, 1996.
- [9] M. Furuse, “Knockout animals and natural mutations as experimental and diagnostic tool for studying tight junction functions in vivo,” pp. 813–819, 2009.
- [10] M. B. Dennis, “Welfare issues of genetically modified animals,” *ILAR Journal*, vol. 43, no. 2, pp. 100–109, 2002.
- [11] M. Rose, “Welfare Phenotyping of Genetically-modified Mice,” pp. 181–186, 2009.
- [12] D. J. Mellor, “Updating Animal Welfare Thinking: Moving beyond the ” Five Freedoms ” towards ” A Life Worth Living ”,” *Animals*, vol. 6, no. 21, pp. 1–20, 2016.
- [13] P. W. Webb, “Effects of Median-Fin Amputation on Fast-Start Performance of Rainbow Trout (*Salmo Gairdneri*),” *The Journal of Experimental Biology*, vol. 68, no. 1, pp. 123–135, 1977. [Online]. Available: <http://jeb.biologists.org/content/68/1/123{%}%5Cnhttp://jeb.biologists.org/content/68/1/123.full.pdf{%}%5Cnhttp://jeb.biologists.org/content/68/1/123.short>
- [14] L. A. Jawad, M. Ibrahim, and B. Waryani, “Incidences of caudal fin malformation in fishes from Jubail City, Saudi Arabia, Arabian Gulf,” *Archives of Polish Fisheries*, vol. 26, no. 1, pp. 65–71, 2018. [Online]. Available: <http://content.sciendo.com/view/journals/aopf/26/1/article-p65.xml>
- [15] C. Best, H. Ikert, D. J. Kostyniuk, P. M. Craig, L. Navarro-Martin, L. Marandel, and J. A. Mennigen, “Epigenetics in teleost fish: From molecular mechanisms to physiological phenotypes,” 2018.
- [16] S. Rajaram, S. Patel, G. K. Uggini, I. Desai, and S. Balakrishnan, “BMP signaling regulates the skeletal and connective tissue differentiation during caudal fin regeneration in sailfin molly (*Poecilia latipinna*),” *Development Growth and Differentiation*, vol. 59, no. 8, pp. 629–638, 2017.
- [17] B. N. Giepmans, S. R. Adams, M. H. Ellisman, and R. Y. Tsien, “The fluorescent toolbox for assessing protein location and function,” pp. 217–224, 2006.

- [18] I. Fabre, “Alternatives to animal testing,” *Bulletin de l’Academie nationale de medecine*, vol. 193, no. 8, pp. 1783–1791, 2009.
- [19] E. S. Allman, J. H. Degnan, and J. A. Rhodes, “Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent,” *Journal of Mathematical Biology*, vol. 62, no. 6, pp. 833–862, 2011.
- [20] H. Philippe, H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain, “Resolving difficult phylogenetic questions: Why more sequences are not enough,” *PLoS Biology*, vol. 9, no. 3, 2011.
- [21] S. Mirarab and T. Warnow, “ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes,” in *Bioinformatics*, vol. 31, no. 12, 2015, pp. i44–i52.
- [22] B. Boussau, G. J. Szollosi, L. Duret, M. Gouy, E. Tannier, and V. Daubin, “Genome-scale coestimation of species and gene trees,” *Genome Research*, vol. 23, no. 2, pp. 323–330, 2013.
- [23] M. Nute, J. Chou, E. K. Molloy, and T. Warnow, “The performance of coalescent-based species tree estimation methods under models of missing data,” *BMC Genomics*, vol. 19, no. Suppl 5, pp. 1–22, 2018. [Online]. Available: <http://dx.doi.org/10.1186/s12864-018-4619-8>
- [24] M. Bogusz and S. Whelan, “Phylogenetic tree estimation with and without alignment: New distance methods and benchmarking,” *Systematic Biology*, vol. 66, no. 2, pp. 218–231, 2017.
- [25] G. Cardona, A. Mir, and F. Rosselló, “The expected value under the Yule model of the squared path-difference distance,” *Applied Mathematics Letters*, vol. 25, no. 12, pp. 2031–2036, 2012.
- [26] M. K. Kuhner and J. Yamato, “Practical performance of tree comparison metrics,” *Systematic Biology*, vol. 64, no. 2, pp. 205–214, 2015.
- [27] T. M. W. Nye, P. Liò, and W. R. Gilks, “A novel algorithm and web-based tool for comparing two alternative phylogenetic trees,” *Bioinformatics*, vol. 22, no. 1, pp. 117–119, 2006.
- [28] D. F. Robinson and L. R. Foulds, “Comparison of phylogenetic trees,” *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.
- [29] D. Penny, L. R. Foulds, and M. D. Hendy, “Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences,” *Nature*, vol. 297, no. 5863, pp. 197–200, 1982.
- [30] Y. Wu, “A practical method for exact computation of subtree prune and regraft distance,” *Bioinformatics*, vol. 25, no. 2, pp. 190–196, 2009.

- [31] A. D. Finden, C. R. , Gordon, “Obtaining common pruned trees,” *J. Classification*, vol. 2, pp. 255–276, 1985.
- [32] W. T. Williams and H. T. Clifford, “On the Comparison of Two Classifications of the Same Set of Elements,” *Taxon*, vol. 20, no. 4, pp. 519–522, 1971. [Online]. Available: <http://www.jstor.org/stable/1218253>
- [33] D. Koufaty and D. T. Marr, “Hyperthreading technology in the netburst microarchitecture,” *IEEE Micro*, vol. 23, no. 2, pp. 56–65, 2003.
- [34] N. H. Qun, Z. I. Khalib, M. N. Warip, M. E. Elobaid, R. Mostafijur, N. A. Zahri, and P. Saad, “Hyper-threading technology: Not a good choice for speeding up CPU-bound code,” in *2016 3rd International Conference on Electronic Design, ICED 2016*, 2017, pp. 578–581.
- [35] V. Nélis, P. M. Yomsi, and L. M. Pinho, “The Variability of Application Execution Times on a Multi-Core Platform,” *Proceedings of the Workshop on Worst-Case Execution Time Analysis (WCET)*, vol. 55, no. 611016, pp. 6:1—6:11, 2016. [Online]. Available: <http://drops.dagstuhl.de/opus/volltexte/2016/6899>
- [36] L. S. Kubatko and J. H. Degnan, “Inconsistency of phylogenetic estimates from concatenated data under coalescence,” *Systematic Biology*, vol. 56, no. 1, pp. 17–24, 2007.
- [37] J. Jansson, C. Shen, and W.-K. Sung, “Improved Algorithms for Constructing Consensus Trees,” *Journal of the ACM*, vol. 63, no. 3, pp. 1–24, 2016. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=2957788.2925985>
- [38] C. Daskalakis and S. Roch, “Species Trees from Gene Trees Despite a High Rate of Lateral Genetic Transfer: A Tight Bound,” *SIAM ACM Symposium on Discrete Algorithms (SODA’16)*, vol. 1508.01962, no. D, pp. 1621–1630, 2016. [Online]. Available: <http://arxiv.org/abs/1508.01962>
- [39] K. J. Van Nimwegen, R. A. Van Soest, J. A. Veltman, M. R. Nelen, G. J. Van Der Wilt, L. E. Vissers, and J. P. Grutters, “Is the 1000 genome as near as we think? A cost analysis of next-generation sequencing,” *Clinical Chemistry*, vol. 62, no. 11, pp. 1458–1464, 2016.
- [40] M. Feuermann, P. Gaudet, H. Mi, S. E. Lewis, and P. D. Thomas, “Large-scale inference of gene function through phylogenetic annotation of Gene Ontology terms: Case study of the apoptosis and autophagy cellular processes,” *Database*, vol. 2016, 2016.
- [41] M. Pawlik and N. Augsten, “RTED : A Robust Algorithm for the Tree Edit Distance,” *The 38th International Conference on Very Large Data Bases*, pp. 334–345, 2011.

- [42] A. J. Aberer, D. Krompass, and A. Stamatakis, “Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice,” *Systematic Biology*, vol. 62, no. 1, pp. 162–166, 2013.
- [43] A. Kawalia, S. Motameny, S. Wonczak, H. Thiele, L. Nieroda, K. Jabbari, S. Borowski, V. Sinha, W. Gunia, U. Lang, V. Achter, and P. Nürnberg, “Leveraging the power of high performance computing for next generation sequencing data analysis: Tricks and twists from a high throughput exome workflow,” *PLoS ONE*, vol. 10, no. 5, 2015.
- [44] U. Manzoor, S. Shahid, and B. Zafar, “A comparative analysis of multiple sequence alignments for biological data,” *Bio-Medical Materials and Engineering*, vol. 26, pp. S1781–S1789, 2015.
- [45] E. K. Molloy and T. Warnow, “To include or not to include : The impact of gene filtering on species tree estimation methods,” *Systematic Biology*, vol. 0, no. September, pp. 1–29, 2017.
- [46] T. N. Theis and H.-S. P. Wong, “The End of Moore’s Law: A New Beginning for Information Technology,” *Computing in Science & Engineering*, vol. 19, no. 2, pp. 41–50, 2017. [Online]. Available: <http://ieeexplore.ieee.org/document/7878935/>
- [47] R. Groisberg, J. Roszik, A. Conley, S. R. Patel, and V. Subbiah, “The Role of Next-Generation Sequencing in Sarcomas: Evolution From Light Microscope to Molecular Microscope,” 2017.
- [48] D. E. Critchlow, D. K. Pearl, and C. Qian, “The triples distance for rooted bifurcating phylogenetic trees,” *Systematic Biology*, vol. 45, no. 3, pp. 323–334, 1996.
- [49] A. Goëffon, J. M. Richer, and J. K. Hao, “Progressive tree neighborhood applied to the maximum parsimony problem,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 1, pp. 136–145, 2008.
- [50] C. C. Ribeiro and D. S. Vianna, “A GRASP/VND heuristic for the phylogeny problem using a new neighborhood structure,” *International Transactions in Operational Research*, vol. 12, no. 3, pp. 325–338, 2005.
- [51] M. Steel and C. Leuenberger, “The optimal rate for resolving a near-polytomy in a phylogeny,” *Journal of Theoretical Biology*, vol. 420, pp. 174–179, 2017.
- [52] A. Kuzniar, R. C. H. J. van Ham, S. Pongor, and J. A. M. Leunissen, “The quest for orthologs: finding the corresponding gene across genomes,” pp. 539–551, 2008.

- [53] G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente, “Metrics for phylogenetic networks i: Generalizations of the robinson-foulds metric,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 1, pp. 46–61, 2009.
- [54] A. Francis, K. T. Huber, V. Moulton, and T. Wu, “Bounds for phylogenetic network space metrics,” pp. 1–20, 2017.
- [55] S. Jiao, L. Hsu, C. M. Hutter, and U. Peters, “The use of imputed values in the meta-analysis of genome-wide association studies,” *Genetic Epidemiology*, vol. 35, no. 7, pp. 597–605, 2011.
- [56] S. R. Browning, “Missing data imputation and haplotype phase inference for genome-wide association studies,” pp. 439–450, 2008.
- [57] A. Korte and A. Farlow, “The advantages and limitations of trait analysis with GWAS: A review,” 2013.
- [58] A. Platt, B. J. Vilhjálmsson, and M. Nordborg, “Conditions under which genome-wide association studies will be positively misleading,” *Genetics*, vol. 186, no. 3, pp. 1045–1052, 2010.
- [59] M. A. Taub, H. Schwender, T. H. Beaty, T. A. Louis, and I. Ruczinski, “Incorporating genotype uncertainties into the genotypic TDT for main effects and gene-environment interactions,” *Genetic epidemiology*, vol. 36, no. 3, pp. 225–234, 2012.
- [60] A. D. Foote, Y. Liu, G. W. Thomas, T. Vinař, J. Alföldi, J. Deng, S. Dugan, C. E. Van Elk, M. E. Hunter, V. Joshi, Z. Khan, C. Kovar, S. L. Lee, K. Lindblad-Toh, A. Mancia, R. Nielsen, X. Qin, J. Qu, B. J. Raney, N. Vijay, J. B. Wolf, M. W. Hahn, D. M. Muzny, K. C. Worley, M. T. P. Gilbert, and R. A. Gibbs, “Convergent evolution of the genomes of marine mammals,” *Nature Genetics*, vol. 47, no. 3, pp. 272–275, 2015.
- [61] I. Y. Kwak and W. Pan, “Gene- and pathway-based association tests for multiple traits with GWAS summary statistics,” *Bioinformatics*, vol. 33, no. 1, pp. 64–71, 2017.
- [62] J. Parker, G. Tsagkogeorga, J. A. Cotton, Y. Liu, P. Provero, E. Stupka, and S. J. Rossiter, “Genome-wide signatures of convergent evolution in echolocating mammals,” *Nature*, vol. 502, no. 7470, pp. 228–231, 2013.
- [63] W. W. Au and K. J. Benoit-Bird, “Automatic gain control in the echolocation system of dolphins,” *Nature*, vol. 423, no. 6942, pp. 861–863, 2003.
- [64] D. R. Ketten, “The marine mammal ear: Specializations for aquatic audition and echolocation,” *The Evolutionary Biology of Hearing*, pp. 717–750, 1992.

- [65] Z. Zou and J. Zhang, “No genome-wide protein sequence convergence for echolocation,” *Molecular Biology and Evolution*, vol. 32, no. 5, pp. 1237–1241, 2015.
- [66] Y. Li, Z. Liu, P. Shi, and J. Zhang, “The hearing gene Prestin unites echolocating bats and whales,” 2010.
- [67] G. W. Thomas and M. W. Hahn, “Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals,” *Molecular Biology and Evolution*, vol. 32, no. 5, pp. 1232–1236, 2015.
- [68] J. J. Wiens, D. S. Moen, and S. Brook, “Missing data and the accuracy of Bayesian phylogenetics,” *Journal of Systematics and Evolution*, vol. 46, no. 3, pp. 307–314, 2008.
- [69] T. A. Castoe, A. P. J. de Koning, H.-M. Kim, W. Gu, B. P. Noonan, G. Naylor, Z. J. Jiang, C. L. Parkinson, and D. D. Pollock, “Evidence for an ancient adaptive episode of convergent molecular evolution,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 22, pp. 8986–8991, 2009.
- [70] E. K. Molloy and T. Warnow, “To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods,” *Systematic Biology*, vol. 67, no. 2, pp. 285–303, 2018.
- [71] D. Fraix-Burnet, “Phylogenetic Tools in Astrophysics,” in *Wiley StatsRef: Statistics Reference Online*, 2017, pp. 1–6. [Online]. Available: <http://doi.wiley.com/10.1002/9781118445112.stat07935>
- [72] T. Onai, N. Irie, and S. Kuratani, “The Evolutionary Origin of the Vertebrate Body Plan: The Problem of Head Segmentation,” *Annual Review of Genomics and Human Genetics*, vol. 15, no. 1, pp. 443–459, 2014. [Online]. Available: <http://www.annualreviews.org/doi/10.1146/annurev-genom-091212-153404>
- [73] N. M. Foley, M. S. Springer, and E. C. Teeling, “Mammal madness: is the mammal tree of life not yet resolved?” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1699, p. 20150140, 2016. [Online]. Available: <http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2015.0140>
- [74] M. E. Alfaro, F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, and L. J. Harmon, “Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 32, pp. 13 410–13 414, 2009. [Online]. Available: <http://www.pnas.org/cgi/doi/10.1073/pnas.0811087106>

- [75] A. W. Shingleton and W. A. Frankino, “New perspectives on the evolution of exaggerated traits,” pp. 100–107, 2013.
- [76] D. W. Davidson, “Size Variability in the Worker Caste of a Social Insect (*Veromessor pergandei* Mayr) as a Function of the Competitive Environment,” *The American Naturalist*, vol. 112, p. 523, 1978. [Online]. Available: <http://www.jstor.org/stable/10.2307/2460120>{%}5Cnpapers2://publication/uuid/C7347C86-6FAE-4268-9DB5-4E35B566F736
- [77] N. Wang, P. A. Hosner, B. Liang, E. L. Braun, and R. T. Kimball, “Historical relationships of three enigmatic phasianid genera (Aves: Galliformes) inferred using phylogenomic and mitogenomic data,” *Molecular Phylogenetics and Evolution*, vol. 109, pp. 217–225, 2017.
- [78] A. Amores, T. Suzuki, Y.-L. Yan, J. Pomeroy, A. Singer, C. Amemiya, and J. H. Postlethwait, “Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish.” *Genome research*, vol. 14, no. 1, pp. 1–10, jan 2004.
- [79] C. Lvque, T. Oberdorff, D. Paugy, M. L. Stiassny, and P. A. Tedesco, “Global diversity of fish (Pisces) in freshwater,” pp. 545–567, 2008.
- [80] D. Booth, E. Poloczanska, J. Donelson, J. Garcia Molinos, and M. Burrows, “Biodiversity and Climate Change in the Oceans,” *Climate Change Impacts on Fisheries and Aquaculture*, 2018.
- [81] R. Betancur-r, R. E. Broughton, E. O. Wiley, K. Carpenter, J. A. Lpez, C. Li, N. I. Holcroft, D. Arcila, M. Sanciangco, J. C. C. Li, F. Zhang, M. a. Campbell, J. a. Ballesteros, A. Roa-varon, S. Willis, W. C. Borden, D. J. Hough, and G. Lu, “The Tree of Life and a New Classification of Bony Fishes,” *PLOS Currents Tree of Life*, vol. 0732988, pp. 1–45, 2013.
- [82] R. Froese and D. Pauly, “FishBase. A Global Information System on Fishes,” 2017. [Online]. Available: <http://www.fishbase.org/home.htm>{%}0Ahttp://www.fishbase.org/
- [83] R. Hovmller, L. Lacey Knowles, and L. S. Kubatko, “Effects of missing data on species tree estimation under the coalescent,” *Molecular Phylogenetics and Evolution*, vol. 69, no. 3, pp. 1057–1062, 2013.
- [84] B. Haubold, “Alignment-free phylogenetics and population genetics,” *Briefings in Bioinformatics*, vol. 15, no. 3, pp. 407–418, 2014.
- [85] K. Nguyen, X. Guo, and Y. Pan, “Multiple Sequences Alignment Algorithms,” in *Multiple Biological Sequence Alignment*. John Wiley & Sons, Inc., 2016, pp. 69–101. [Online]. Available: <http://dx.doi.org/10.1002/9781119273769.ch5>

- [86] K. Katoh and D. M. Standley, “MAFFT multiple sequence alignment software version 7: Improvements in performance and usability,” *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 2013.
- [87] F. Sievers and D. G. Higgins, “Clustal omega, accurate alignment of very large numbers of sequences,” *Methods in Molecular Biology*, vol. 1079, pp. 105–116, 2014.
- [88] S. Mirarab, N. Nguyen, and T. Warnow, “SEPP: SATé-Enabled Phylogenetic Placement,” in *Biocomputing 2012*, 2011, pp. 247–258.
- [89] T. Lassmann and E. L. L. Sonnhammer, “Kalign—an accurate and fast multiple sequence alignment algorithm.” *BMC bioinformatics*, vol. 6, p. 298, 2005.
- [90] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, “Mauve: Multiple alignment of conserved genomic sequence with rearrangements,” *Genome Research*, vol. 14, no. 7, pp. 1394–1403, 2004.
- [91] S. Shehab, S. Abdulah, and A. E. Keshk, “A Survey of the State-of-the-Art Parallel Multiple Sequence Alignment Algorithms on Multicore Systems,” no. x, pp. 1–16, 2018. [Online]. Available: <http://arxiv.org/abs/1805.12223>
- [92] A. Stamatakis, “RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies,” *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, 2014.
- [93] J. C. Wilgenbusch and D. Swofford, “Inferring Evolutionary Trees with PAUP*,” in *Current Protocols in Bioinformatics*, 2003.
- [94] L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, “IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies,” *Molecular Biology and Evolution*, vol. 32, no. 1, pp. 268–274, 2015.
- [95] J. P. Huelsenbeck and F. Ronquist, “MR BAYES: Bayesian inference of phylogenetic trees,” *Bioinformatics*, vol. 17, pp. 754–755, 2001.
- [96] J. A. Lees, M. Kendall, J. Parkhill, C. Colijn, S. D. Bentley, and S. R. Harris, “Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study,” *Wellcome Open Research*, vol. 3, no. May, p. 33, 2018. [Online]. Available: <https://wellcomeopenresearch.org/articles/3-33/v1>
- [97] M. N. Price, P. S. Dehal, and A. P. Arkin, “FastTree 2—approximately maximum-likelihood trees for large alignments.” *PloS one*, vol. 5, no. 3, p. e9490, jan 2010. [Online].

Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2835736&tool=pmcentrez&rendertype=abstract>

- [98] M. D. Hellwig and N. Daniels, “Taxon Focused Tree Distance Estimation (TDEFT),” *ISMB*, 2019.
- [99] A. D. Finden, C. R. , Gordon, “Obtaining common pruned trees,” *J. Classification*, vol. 2, pp. 255–276, 1985.
- [100] D. Robinson and L. Foulds, “Comparison of weighted labelled trees,” *Lecture Notes in Mathematics*, vol. 748, pp. 119–126, 1979. [Online]. Available: <http://dx.doi.org/10.1007/BFb0102690>{%}5Cn<http://www.springerlink.com/content/j1w15867137r8753/>
- [101] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.
- [102] H. Huang, J. Sukumaran, S. A. Smith, and L. L. Knowles, “Cause of gene tree discord? Distinguishing incomplete lineage sorting and lateral gene transfer in phylogenetics,” *PeerJ Preprints*, no. i, 2017.
- [103] S. Schaack, C. Gilbert, and C. Feschotte, “Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution,” pp. 537–546, 2010.
- [104] P. J. Keeling, “Functional and ecological impacts of horizontal gene transfer in eukaryotes,” pp. 613–619, 2009.
- [105] G. L. Wallau, M. F. Ortiz, and E. L. S. Loreto, “Horizontal transposon transfer in eukarya: Detection, bias, and perspectives,” *Genome Biology and Evolution*, vol. 4, no. 8, pp. 689–699, 2012.
- [106] J. Yang and T. Warnow, “Fast and accurate methods for phylogenomic analyses,” *BMC Bioinformatics*, vol. 12, no. SUPPL. 9, 2011.
- [107] S. Song, L. Liu, S. V. Edwards, and S. Wu, “Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 37, pp. 14 942–14 947, 2012.
- [108] L. Nakhleh, “Computational approaches to species phylogeny inference and gene tree reconciliation,” pp. 719–728, 2013.
- [109] D. E. Hunt, J. G. Purser, G. A. Thomas, J. M. Cobcroft, and N. J. F. Rawlinson, “Caudal fin aspect ratio as a predictor of maximum swimming speed,” in *Australian Society for Fish Biology-Oceania Chondrichthyan Society Conference*, 2016, p. .

- [110] R. Derynck and Y. E. Zhang, “Smad-dependent and Smad-independent pathways in TGF-,” *Nature*, vol. 425, pp. 577–584, 2003.
- [111] M. Schebesta, C. L. Lien, F. B. Engel, and M. T. Keating, “Transcriptional profiling of caudal fin regeneration in zebrafish,” *TheScientificWorldJournal*, vol. 6, no. SUPPL.1, pp. 38–54, 2006.
- [112] C. V. Esguerra, L. Nelles, L. Vermeire, A. Ibrahimi, A. D. Crawford, R. Derua, E. Janssens, E. Waelkens, P. Carmeliet, D. Collen, and D. Huylebroeck, “Ttrap is an essential modulator of Smad3-dependent Nodal signaling during zebrafish gastrulation and left-right axis determination,” *Development*, vol. 134, no. 24, pp. 4381–4393, 2007. [Online]. Available: <http://dev.biologists.org/cgi/doi/10.1242/dev.000026>
- [113] J. A. Lister, B. M. Lane, A. Nguyen, and K. Lunney, “Embryonic expression of zebrafish MiT family genes Tfe3b, Tfeb, and Tfec,” *Developmental Dynamics*, vol. 240, no. 11, pp. 2529–2538, 2011.
- [114] P. Hu, J. Shangguan, and L. Zhang, “Downregulation of NUF2 inhibits tumor growth and induces apoptosis by regulating lncRNA AF339813,” *International Journal of Clinical and Experimental Pathology*, vol. 8, no. 3, pp. 2638–2648, 2015.
- [115] A. Benini, F. Cignarella, L. Calvarini, S. Mantovanelli, E. Giacomuzzi, D. Zizioli, and G. Borsani, “Slc7a6os gene plays a critical role in defined areas of the developing CNS in Zebrafish,” *PLoS ONE*, vol. 10, no. 3, 2015.
- [116] L. B. Miles, C. Darido, J. Kaslin, J. K. Heath, S. M. Jane, and S. Dworkin, “Mis-expression of grainyhead-like transcription factors in zebrafish leads to defects in enveloping layer (EVL) integrity, cellular morphogenesis and axial extension,” *Scientific Reports*, vol. 7, no. 1, 2017.
- [117] Y. Xie, D. Yang, Q. He, and Z. Songyang, “Zebrafish as a model system to study the physiological function of telomeric protein TPP1,” *PLoS ONE*, vol. 6, no. 2, 2011.
- [118] A. Amsterdam, R. M. Nissen, Z. Sun, E. C. Swindell, S. Farrington, and N. Hopkins, “Identification of 315 genes essential for early zebrafish development,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 35, pp. 12 792–12 797, 2004. [Online]. Available: <http://www.pnas.org/cgi/doi/10.1073/pnas.0403929101>
- [119] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law, “Comparison of next-generation sequencing systems,” 2012.
- [120] B. Thisse and C. Thisse, “Fast Release Clones: A High Throughput Expression Analysis,” 2004. [Online]. Available: <http://zfin.org/ZDB-PUB-040907-1>

- [121] P. Coutinho, M. J. Parsons, K. A. Thomas, E. M. A. Hirst, L. Saúde, I. Campos, P. H. Williams, and D. L. Stemple, “Differential requirements for COPI transport during vertebrate early development,” *Developmental Cell*, vol. 7, no. 4, pp. 547–558, 2004.
- [122] A. Barbieri, S. Carra, P. De Blasio, F. Cotelli, and I. Biunno, “Sel1l knock-down negatively influences zebrafish embryos endothelium,” *Journal of Cellular Physiology*, vol. 233, no. 7, pp. 5396–5404, 2018.
- [123] M. Brand, C. P. Heisenberg, R. M. Warga, F. Pelegri, R. O. Karlstrom, D. Beuchle, a. Picker, Y. J. Jiang, M. Furutani-Seiki, F. J. van Eeden, M. Granato, P. Haffter, M. Hammerschmidt, D. a. Kane, R. N. Kelsh, M. C. Mullins, J. Odenthal, and C. Nüsslein-Volhard, “Mutations affecting development of the midline and general body shape during zebrafish embryogenesis.” *Development (Cambridge, England)*, vol. 123, no. January 1997, pp. 129–142, 1996.
- [124] Y. Zhang and A. V. Alekseyenko, “Phylogenetic inference using alignment-free methods for applications in microbial community surveys using 16s rRNA gene,” *PLoS ONE*, vol. 12, no. 11, 2017.
- [125] P. V. Asharani, K. Keupp, O. Semler, W. Wang, Y. Li, H. Thiele, G. Yigit, E. Pohl, J. Becker, P. Frommolt, C. Sonntag, J. Altmüller, K. Zimmermann, D. S. Greenspan, N. A. Akarsu, C. Netzer, E. Schönau, R. Wirth, M. Hammerschmidt, P. Nürnberg, B. Wollnik, and T. J. Carney, “Attenuated BMP1 function compromises osteogenesis, leading to bone fragility in humans and Zebrafish,” *American Journal of Human Genetics*, vol. 90, no. 4, pp. 661–674, 2012.
- [126] X. He, W. Zhang, C. Yan, F. Nie, C. Li, X. Liu, C. Fei, S. Li, X. Song, Y. Jia, R. Zeng, D. Wu, W. Pan, X. Hao, and L. Li, “Chemical biology reveals CARF as a positive regulator of canonical Wnt signaling by promoting TCF/ β -catenin transcriptional activity,” *Cell Discovery*, vol. 3, 2017.
- [127] R. Bammens, N. Mehta, V. Race, F. Foulquier, J. Jaeken, M. Tiemeyer, R. Steet, G. Matthijs, and H. Flanagan-Steet, “Abnormal cartilage development and altered N-glycosylation in Tmem165-deficient zebrafish mirrors the phenotypes associated with TMEM165-CDG,” *Glycobiology*, vol. 25, no. 6, pp. 669–682, 2014.
- [128] B. L. Arduini, K. M. Bosse, and P. D. Henion, “Genetic ablation of neural crest cell diversification,” *Development*, vol. 136, no. 12, pp. 1987–1994, 2009. [Online]. Available: <http://dev.biologists.org/cgi/doi/10.1242/dev.033209>
- [129] S. C. Neuhauss, L. Solnica-Krezel, A. F. Schier, F. Zwartkruis, D. L. Stemple, J. Malicki, S. Abdelilah, D. Y. Stainier, and W. Driever,

- “Mutations affecting craniofacial development in zebrafish.” *Development (Cambridge, England)*, vol. 123, pp. 357–67, 1996. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/9007255>
- [130] T. F. Schilling, T. Piotrowski, H. Grandel, M. Brand, C. P. Heisenberg, Y. J. Jiang, D. Beuchle, M. Hammerschmidt, D. A. Kane, M. C. Mullins, F. J. van Eeden, R. N. Kelsh, M. Furutani-Seiki, M. Granato, P. Haffter, J. Odenthal, R. M. Warga, T. Trowe, and C. Nüsslein-Volhard, “Jaw and branchial arch mutants in zebrafish I: branchial arches.” *Development (Cambridge, England)*, vol. 123, pp. 329–44, 1996. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/9007253>
- [131] G. Gestri, R. J. Osborne, A. W. Wyatt, D. Gerrelli, S. Gribble, H. Stewart, A. Fryer, D. J. Bunyan, K. Prescott, J. R. O. Collin, T. Fitzgerald, D. Robinson, N. P. Carter, S. W. Wilson, and N. K. Ragge, “Reduced TFAP2A function causes variable optic fissure closure and retinal defects and sensitizes eye development to mutations in other morphogenetic regulators,” *Human Genetics*, vol. 126, no. 6, pp. 791–803, 2009.
- [132] P. J. Lyons, M. R. Sapio, and L. D. Fricker, “Zebrafish cytosolic carboxypeptidases 1 and 5 are essential for embryonic development,” *Journal of Biological Chemistry*, vol. 288, no. 42, pp. 30 454–30 462, 2013.
- [133] R. Landel, “Feedback: Letters from our readers,” *Genetics*, vol. 94, no. 1, pp. 203–215, 1980.