Multilocus Phylogenetic Analysis with Gene Tree Clustering *

Ruriko Yoshida[1], Kenji Fukumizu[2], and Chrysafis Vogiatzis[†3]

[1]*Department of Statistics,, University of Kentucky,, Lexington, KY, USA*
[2]*Department of Industrial & Manufacturing Engineering,, North Dakota State University,, Fargo, ND, USA*
[3]*Department of Industrial & Manufacturing Engineering,, North Dakota State University,, Fargo, ND, USA*

July 17, 2016

Abstract

Both theoretical and empirical evidence seem to indicate that phylogenetic trees of different genes (loci) do not display precisely matched topologies. This phylogenetic incongruence is mainly attributed to the reticulated evolutionary history of most species, due to evolutionary events such as meiotic sexual recombination in eukaryotes, or horizontal transfers of genetic material in prokaryotes. Nonetheless, most genes do display topologically related phylogenies; this implies they form cohesive subsets (clusters). In this work, we discuss clustering of gene trees, focusing on the normalized cut (Ncut) framework as a suitable clustering method for phylogenetics. We proceed to show that this framework is indeed efficient and statistically accurate when obtaining clusters on the set of gene trees based on the geodesic distance between them over the Billera-Holmes-Vogtmann (BHV) tree space.

[*] JSPS KAKENHI 26540016 and ND EPSCoR NSF 1355466

[†] E-mail: `chrysafis.vogiatzis@ndsu.edu`; Corresponding author

We then conduct a computational study on the performance of different clustering methods, with and without preprocessing, under different distance metrics, and using a series of dimension reduction techniques. Our results with simulated data reveal that the Ncut framework accurately clusters the set of gene trees, given a species tree under the coalescent process. We further depict the success of our framework by comparing its performance to other clustering techniques, including k-means and hierarchical clustering. The main computational results can be summarized to the better performance of the Ncut framework even without dimension reduction, the similar performance portrayed by Ncut and k-means under most dimension reduction schemes, the worse performance of hierarchical clustering to accurately capture clusters, as well as the significantly better performance of the neighbor-joining method with the p-distance (NJp), as compared to the well-studied maximum-likelihood estimation method (MLE). Supplementary material, all codes, and the data used in this work are freely available at `http://polytopes.net/research/cluster/` online.

Keywords. Phylogenetics; Normalized Cut; Clustering

1 Introduction

During this last decade, the field of phylogenetics has been undergoing a gradual shift away from the notion of the strictly bifurcating, completely resolved species trees, to a recognition that species are containers of allelic variation for each gene. It is very well established that differences in lineage sorting due to genetic drift lead to differences in phylogenetic tree topologies (Maddison, 1997). Gene flow in ancestral populations and independent lineage sorting of polymorphisms is fully expected to generate topological conflicts between gene trees in reticulating (e.g., sexually recombining) species (Taylor *et al.*, 2000; Huson *et al.*, 2005; Weisrock *et al.*, 2006). Both extant and ancestral species could exhibit this phenomenon, so ancestral species should not be regarded as node points in a fully resolved bifurcating tree, but instead they can be thought of as spatiotemporal clouds of individual genotypes with all their inherent allelism. Thus, a central issue in systematic biology is the reconstruction of populations and species from numerous gene trees with varying levels of discordance (Brito and Edwards, 2009; Edwards, 2009). While there has been a well-established understanding of the discordant phylogenetic relationships that can exist among independent gene trees drawn from a common species tree (Pamilo and Nei, 1988; Takahata, 1989; Maddison, 1997; Bollback and Huelsenbeck, 2009), phylogenetic studies have only recently begun to shift away from single gene or concatenated gene estimates of phylogeny towards these multilocus approaches (e.g. Carling and Brumfield (2008); Yu *et al.* (2011); Betancur *et al.* (2013); Heled and Drummond (2011); Thompson and Kubatko (2013)). Initially, the study of phylogenetics tended to focus on individual protein sequences (Neyman, 1971). However, with the availability of more and more sequencing data, it has become imperative to consider more loci in our studies, which will lead to significantly less biased phylogenetic inferences, since uncertainty, along with missing data, is mitigated over multiple loci.

There are several methods to conduct multilocus phylogenetic analysis (Bininda-Emonds *et al.*, 2002; Liu *et al.*, 2009). Most of them infer the best fit tree from the entire data set. However this "averaging scheme" over multiple loci might not work to handle biological processes such as *horizontal gene transfer* (Jeffroy *et al.*, 2006). It has been documented that it is common for the genome to have multiple different evolutionary histories (Leigh *et al.*, 2011), which in turn leads to a significant reduction in the correlation among gene trees. There exist numerous processes that can reduce this correlation. As an example, negative or balancing selection on a particular locus is expected to increase the probability that ancestral gene copies are maintained through speciation events (Takahata and Nei, 1990). Furthermore, horizontal transfer shuffles divergent genes among different species (Maddison, 1997). Correlation may also be reduced by naive sampling of loci for analysis. For example, paralogous gene copies will result in a gene tree that conflates gene duplication with speciation. Similarly, sampled sequence data that span one or more recombination events will yield "gene trees" that are hybrids of two or more genealogical histories (Posada and Crandall, 2002). These non-coalescent processes can strongly influence phylogenetic inference (Martin and Burg, 2002; Posada and Crandall, 2002; Edwards, 2009) and are often misleading. In addition, Rivera *et al.* (1998) showed that an analysis of complete genomes indicated a massive prokaryotic gene transfer (or transfers) preceding the formation of the eukaryotic cell, arguing that there is significant genomic evidence for more than one distinct class of genes. As for a specific example, Gori *et al.* (2015) showed that there are three clusters in the data set consisting of 344 curated orthologous sets of genes from 18 ascomycetous yeast species from Hess and Goldman (2011). These examples suggest that the distribution of gene trees may be more accurately modeled as a mixture of a number of more fundamental distributions. In order to find a mixture structure in distributions of gene trees, the first step is to accurately identify the components in the mixture. This is the main reason why in this work we focus on the problem of *clustering* gene trees over the "tree space", in order to detect subsets (clusters) of maximum similarity.

Many researchers take an approach to apply a likelihood based method, such as the *maximum likelihood estimator* (MLE) or *Bayesian inference* on the *concatenated alignment* from gene alignments in order to reconstruct the species tree. However, Roch and Steel (2015) showed that if we apply a likelihood based method on the concatenated alignment from gene alignments, then the resulting trees might be statistically inconsistent because some gene trees are significantly incongruent from the species tree due to incomplete lineage sorting and horizontal gene transfer, among other reasons. More precisely, they showed that under the multi-species coalescent with a standard site substitution model, such as the general time reversible (GTR) model (Tavare, 1986), the MLE method on sequence data that has been concatenated across genes was shown to be a statistically inconsistent estimator of the species, under the mild assumption that all sites have evolved independently and identically on a fixed tree.

Here we focus on a *non-parametric* approach in order to detect the existence and extent of any significant incongruence within a given data set, without relying on any prior assumptions about its biological basis. There are several non-parametric methods in the form of statistical tests of incongruence such as Holmes (2005); Haws *et al.* (2012); Weyenberg *et al.* (2014). Typically, statistical analysis on phylogenetic trees is conducted by mapping each tree to a vector in $\mathbb{R}^d$, $d \in \mathbb{N}$: this is referred to as a *dissimilarity map* (e.g., Holmes (2005); Haws *et al.* (2012); Weyenberg *et al.* (2014); Gori *et al.* (2015)). Given any tree T of n leaves with branch length information, one may produce a corresponding *distance matrix*, $D(T)$. This distance matrix is an $n \times n$ symmetric matrix of non-negative real numbers, with elements corresponding to the sum of the branch lengths between pairs of leaves in the tree. To calculate $D_{(ij)}(T)$, one simply determines which edges of the tree form the path from a leaf i to a leaf j, and then sums the lengths of these branches. Since $D(T)$ is symmetric and has zeros on the diagonal, the upper-triangular portion of the matrix contains all of the unique information required for analysis. We can vectorize a tree T by enumerating this unique portion of the distance matrix,

$$v_D(T) := (D_{12}(T), D_{13}(T), \ldots, D_{23}(T), \ldots, D_{n-1,n}(T)),$$

which is called the *dissimilarity map* of a tree T and is a vector in $\mathbb{R}^{\binom{n}{2}}$.

However, the space of phylogenetic trees with n leaves is not a Euclidean space. In fact, it is represented as the union of lower dimensional polyhedral cones in $\mathbb{R}^{\binom{n}{2}}$. Billera *et al.* (2001) introduced a continuous space which explicitly models the set of rooted phylogenetic trees with edge lengths on a fixed set of leaves. Although the Billera-Holmes-Vogtmann (BHV) tree space is not Euclidean, it is non-positively curved, and thus any two points are connected by a unique shortest path through the space, called a *geodesic*. In this computational study, we show, among other things, that using the BHV tree space can help produce more statistically accurate results.

In a relevant study by Gori *et al.* (2015), the above observations have led the authors to investigate different clustering techniques under different distance spaces. More specifically, the Euclidean, geodesic in the BHV tree space, and Robinson-Foulds distance metrics are investigated, while for clustering the authors put to the test hierarchical clustering, spectral clustering, k-means, and k-medoids. They then proceed to use the best obtained setup (spectral clustering with a combination of geodesic and Euclidean distances) to further analyze real-life genome data.

In contrast, we propose the use of the framework of Normalized cut (Ncut), introduced by Shi and Malik (2000); this framework is directly applicable to the distance matrix given by the geodesics in the BHV tree space. The Ncut has been proposed and applied very successfully in the field of image segmentation, but has not been thoroughly investigated in relation to phylogenetics. This paper aims to present the differences in the performance of the normalized cut framework, k-means, and hierarchical clustering in the Euclidean and the BHV tree space, and under different dimension reduction approaches using both simulated and real datasets. Another significant difference between the work of Gori *et al.* (2015) and ours is that we also consider the uncertainty of the tree reconstruction into the analysis by considering the Maximum Likelihood Estimator (MLE) under several different evolutionary models and the Neighbor-Joining method with the p-distance (NJp) while Gori *et al.* (2015) assumed that the set of true gene trees is readily available as their input.

The manuscript is outlined as follows. Section 2 offers a basic review of the

BHV space, the normalized cut framework, and different dimension reduction techniques for the interested reader. In Section 3, we present our computational study and the results obtained using simulated and real datasets. Last, Section 4 discusses these results with a focus on our main computational analysis, while also summarizing our work and offering insight into possible future directions.

2 Fundamentals

Herein, we present a comparative study of different methods for multilocus phylogenetic analysis using gene tree clustering based on the distance matrix obtained by the geodesic distances between tree pairs over the BHV space. The methods we compare are the normalized cut framework, based on the seminal contribution by Shi and Malik (2000), k-means (e.g., Hartigan, 1975), and hierarchical clustering (e.g., (Everitt *et al.*, 2011); interested readers are also referred to Maimon and Rokach (2005) for an excellent overview).

Furthermore, we investigate how dimension reduction methods can be applied in order to extract a lower dimensional structure before clustering and whether that affects the solution quality compared to applying our clustering methods directly upon the original distance matrix. It should be noted that this reduction can also help with regards to data visualization purposes. For dimension reduction, kernel principal component analysis (KPCA, Schölkopf *et al.* (1998)) and t-stochastic neighborhood embedding (t-SNE, van der Maaten and Hinton (2008)) are employed among many other methods, based on our preliminary experiments. Hereafter, we refer to the direct application of clustering to a distance matrix as *direct*, and the above two approaches for dimension reduction as *KPCA* and *t-SNE*, respectively.

We now proceed to offer some basics on the BHV tree space, the normalized cut framework, and the different dimension reduction techniques used. For more details on the methods used, we refer the interested reader to the Supplementary Material available at `http://polytopes.net/research/cluster/`.

2.1 Billera-Holmes-Vogtmann Tree Space

Billera *et al.* (2001) introduced a continuous space that models the set of rooted phylogenetic trees with edge lengths on a fixed set of leaves. Note here that unrooted trees can be accommodated by using either the Ferras transformation, or by designating arbitrarily a leaf node as the root. It is known that in the *Billera-Holmes-Vogtmann (BHV) tree space* any two points are connected by a geodesic, and the distance between two trees is defined as the *length of the geodesic* connecting them.

Consider a rooted tree with n leaves. Such a tree has at most $2n - 2$ edges; there are n terminal edges, which are connected to leaves, and as many as $n - 2$ internal edges. The maximum number of edges is achieved when the tree is binary, but the number of edges can be lower if the tree contains any polytomies. With each distinct tree topology, we associate a Euclidean *orthant* of dimension equal to the number of edges that the topology possesses. Here, we may regard an orthant to be the subset of $\mathbb{R}^d$ with all coordinates being non-negative. For each topology, the orthant coordinates correspond to the edge lengths in the tree.

In the following discussion regarding the BHV space, we can ignore the terminal edge lengths (this does not mean that we ignore the terminal edge lengths in the computation of the geodesic distance), and focus primarily on the portion of each orthant that describes the internal edges. This is justified since all tree topologies have the same set of n terminal leaves, and each of these leaves is associated with a single terminal edge. Therefore the orthant coordinates associated with the terminal edges are of less interest than those of internal nodes. First, we focus on computing the geodesic distance between two trees on the BHV tree space and then we include the terminal branch lengths for calculating the overall geodesic distance between two trees by taking the difference between each terminal branch length. An interested reader can see Owen and Provan (2011) for details in the computation of the geodesic distances.

Recall, at this point, that this space has at most $n-2$ dimensions. Since each of the coordinates in a simplified orthant corresponds to an internal edge length, the orthant boundaries (where at least one coordinate is zero) represent trees with collapsed internal edges. These points can be thought of as trees with slightly different – but closely related – topologies. The BHV space is constructed by noting that the boundary trees from two different orthants may describe the same polytomic topology. With this insight, we may set about constructing the space by grafting orthant boundaries together when the trees they represent coincide.

As each orthant is locally viewed as a Euclidean space, the shortest path between two points within a single orthant is a straight line. The difficulty comes in establishing which sequence of orthants joining the two topologies will contain the geodesic. In the case of four leaves, this could be easily done using a brute-force search, but we cannot hope to do so with larger trees. Owen and Provan (2011) present a quartic-time algorithm (in the number of leaves n) for finding the geodesic path between any two points in the space. Once the geodesic is known, computing its length – and thus the distance between the trees – is a simple feat.

2.2 Clustering

Given a set of gene trees for the species in analysis, a clustering algorithm is applied based on the distance matrix containing the geodesic distances in the BHV tree space. As an alternative, dimension reduction may be applied before the clustering when directly applying the clustering techniques proves unfruitful.

There are many standard clustering methods ranging from non-hierarchical clustering, such as k-means, to hierarchical clustering methods. This paper focuses on three methods: normalized cut, k-means, and hierarchical clustering (with average linkage). The k-means method is the most standard non-hierarchical clustering method in the literature, and has been extensively used in a variety of applications. Note, however, that with BHV geodesic distances, direct application of the k-means method would be not as accurate, as updating the "centroids" (a step that is required in the method) is problematic due to its "stickiness" to the boundaries of the treespace (Miller *et al.*, 2015). As a linkage method for hierarchical clustering we use average linkage, since this is traditionally applied to general distance or dissimilarity measures. Note that another popular choice, Ward's method, lacks variance interpretation for a non-Euclidean distance matrix.

From the clustering methods presented in this computational study, the normalized cut framework (Shi and Malik, 2000) has been successfully applied to numerous applications, including image segmentation (Shi and Malik, 2000; Carballido-Gamio *et al.*, 2004; Yao *et al.*, 2012), biology (Xing and Karp, 2001; Higham *et al.*, 2007), and social networks (Newman, 2013). It has been used in the past also in phylogenetics (Abascal and Valencia, 2002; Chatterji *et al.*, 2008), however its performance when distance metrics vary is still largely uninvestigated. The Ncut framework can be employed for clustering even when only a similarity or dissimilarity matrix is available; that is, the coordinates of the original data points are not necessary. To properly apply the normalized cut framework in a clustering setting the only required input is the set of data points (each represented by a node in an undirected graph) forming node set V, and a set of weights of similarity between them (the edge set, E, of the graph, with w_e representing the similarity, for $e \in E$). Then, the normalized cut framework aims to detect a bipartition of the node set of the graph in two node sets, $(S, \bar{S})$, such that the objective function

$$NCut(S, \bar{S}) = \frac{cut(S, \bar{S})}{assoc(S, V)} + \frac{cut(S, \bar{S})}{assoc(\bar{S}, V)} \tag{1}$$

is minimized with $S \cup \bar{S} = V$ and $S \cap \bar{S} = \emptyset$. We refer to the summation of all weights of the edges within the cluster as the *association* of the cluster ($assoc(S, V) = \sum_{e \in (S \times S) \cap E} w_e$), and the summation of all edges with exactly one endpoint in each cluster $(S, \bar{S})$ as the *cut* ($cut(S, \bar{S}) = \sum_{e \in (S \times \bar{S}) \cap E} w_e$). If we want more than two clusters, one of the clusters is partitioned further by the same criterion.

More recently, the problem has been studied by Hochbaum (2010) and Hochbaum (2013), where normalized cut variants are discussed, with some of them being shown to be solvable in polynomial time. Among them, of interest to the clustering community would be the *"normalized cut"* problem of Sharon *et al.* (2006), which is nothing more but a single version of the original normalized cut criterion shown in (1) and the *ratio regions* problem (Cox *et al.*, 1996). In Hochbaum (2010) both the ratio regions and "normalized cut" problems were shown to be poly-time solvable. In addition, the normalized cut is known to be solved approximately (with typically good performance) as a generalized eigenproblem, which admits a straightforward and easy to implement solution. Using this relaxation as an eigenproblem, the Ncut is similar to solving a spectral clustering problem. Spectral clustering is not new in phylogenetics, and has been studied in the past (Chen *et al.*, 2007; Zhang *et al.*, 2011). A popular method for spectral clustering uses k-means after obtaining a low-dimensional data representation by the spectral method (see Gori *et al.* (2015)). Since from our preliminary experiments the spectral method does not necessarily provide low-dimensional plots whose shape is suitable for k-means clustering, we chose the Ncut method, which uses a graph-based criterion to partition data.

As mentioned earlier, the only necessary information for the Ncut to work is the similarity/dissimilarity information for all data points. As in our case we only have access to a properly constructed distance matrix, we convert the distance between two data points to a similarity w_{ij} by computing $w_{ij} = e^{-\frac{1}{2\sigma^2}D_{ij}^2}$, where D_{ij} is the distance matrix element, and $\sigma = 1.2 \times m$, where $m = \mathrm{median}\{D_{ij} \mid i \neq j\}$. The median has been used as a heuristically good approximate solution for the parameter in Gaussian kernel for kernel methods (Gretton *et al.*, 2005), and the constant 1.2 was found from the preliminary experiments to be a generally good choice.

2.3 Dimension reduction

Optionally, before clustering, a low-dimensional expression of gene trees may be extracted from the distance matrix. Among various dimension reduction methods, kernel principal component analysis (KPCA, Schölkopf *et al.*, 1998) and t-stochastic neighborhood embedding (t-SNE, van der Maaten and Hinton, 2008) are chosen for our analysis by preliminary experiments (we also applied spectral methods and Isomap, but the results were less favorable than KPCA and t-SNE; for a full set of computational results, we refer the interested reader to the Supplementary Material). Three-dimensional expressions of the data were extracted using the above methods, when of course applied, and the clustering methods were then applied to the Euclidean distance matrix among the three-dimensional data points.

KPCA is a nonlinear extension of the standard principal component analysis (PCA); it applies PCA to feature vectors, which are given by a nonlinear mapping of the original data to the feature space. The nonlinear map is defined by a *positive definite kernel*, and the feature space is possibly an infinite dimensional Hilbert space provided implicitly by the positive definite kernel. KPCA gives nonlinear functions $f_1, \ldots f_d$ of data points $(X_i)_{i=1}^N$ so that $(f_1(X_i), \ldots, f_d(X_i))_{i=1}^N$ can serve as a d-dimensional representation of data. The analysis of this paper uses Gaussian kernel $k(X_i, X_j) = \exp(-\frac{1}{2\sigma^2}D_{ij}^2)$ where D_{ij} is the distance matrix of the gene trees[1].

Last, t-SNE is a method for low-dimensional expression or visualization of high-dimensional data; it typically provides us with a two- or three-dimensional expression. Given $(X_i)_{i=1}^N$ in a high-dimensional space, t-SNE first computes a probability p_{ij} based on the distance matrix so that a high

[1] While this kernel with an arbitrary distance matrix D is not necessarily positive definite, in our analysis the Gram matrices $k(X_i, X_j)$ created by the given data were positive definite.

probability corresponds to higher similarity of X_i and X_j. The method then provides a low-dimensional expression $(Y_i)_{i=1}^N$ in such a way that the set of probabilities q_{ij}, defined similarly for a pair (Y_i, Y_j), is close in value to p_{ij}. The points $(Y_i)_{i=1}^N$ are found with numerical optimization to minimize the Kullback-Leibler divergence between the sets of probabilities (p_{ij}) and (q_{ij}) (see Supplementary Material for more details). In our experiments, a Matlab implementation by van der Maaten (lvdmaaten.github.io/tsne/) is used. The perplexity parameter, which provides a way to determine local bandwidth parameters, is set to 30 (as is the default value in the software) in our experiments.

3 Results

We conducted numerical experiments with simulated datasets and a genome dataset. All our simulated datasets were generated using the software `Mesquite` (Maddison and Maddison, 2009). We first demonstrate that the normalized cut framework accurately clusters the set of gene trees given by a species tree under the coalescent process. Then, we proceed to compare different dimension reduction schemes and their performance as compared to clustering via normalized cut directly on the original tree space. Last, we compare k-means and hierarchical clustering to our proposed approach. Two main observations throughout our obtained results are that hierarchical clustering is not effective in recognizing clusters (as opposed to normalized cut and k-means), and that the frameworks perform better on the gene trees reconstructed via the neighbor-joining (NJ) method (Saitou and Nei, 1987) than those reconstructed via the MLE under evolutionary models.

The experimental design for the genome dataset in Subsection 3.2 is as follows. The three clustering methods were first applied to a genome-wide dataset on coelacanths, lungfishes, and tetrapods from Amemiya *et al.* (2013) and Liang *et al.* (2013), where it was observed that there exist two reliable clusters in their 1290 genes. Based on the datasets, we reconstructed the consensus trees using NJ trees with bootstrap confidence for the clusters ≥ 0.95 (see Subsection 3.2 for more details). We performed numerical experiments using the Euclidean space and the BHV tree space, and compared the accuracy of the two spaces for the goal of recognizing more statistically accurate clusters. Last, we compared different preprocessing dimension reduction schemes for each and every one of the spaces, and the clustering techniques.

Overall, we obtained consistent results with both the normalized cut and the k-means frameworks on the consensus trees obtained. The consensus tree from one cluster (of 858 gene trees with the direct application of the normalized cut, of 761 gene trees with the normalized cut after applying KPCA, and of 817 gene trees with t-NSE normalized cut) supports the view of Fritzsch (1987) and Gorr *et al.* (1991) that claims that coelacanths are most closely related to the tetrapods; furthermore, the consensus tree constructed from the other cluster (of 322 gene trees with the direct Ncut algorithm, of 320 gene trees with the KPCA Ncut algorithm, and of 463 gene trees with the t-NSE Ncut) supports the view of Takezaki *et al.* (2004), that is, the coelacanth, lungfish, and tetrapod lineages diverged within a very short time interval and that their relationships may represent an irresolvable trichotomy. We now proceed to describe our results in more detail on both simulated datasets (see Subsection 3.1) and coelacanths, lungfishes, and tetrapods (see Subsection 3.2).

3.1 Simulated data sets

The simulated data is generated as follows. We first fixed the number of species as ten ($n = 10$), the population size $N_e = 10,000$, and set the species depth as $c \cdot N_e$ where $c \in \{0.6, 0.7, 0.8, 0.9, 1, 1.2, 1.4, 1.6, 1.8, 2\}$. Then, for each species depth $c \cdot N_e$, we generated two species trees from the Yule process (with parameters being the default setting of the software `Mesquite` (Maddison and Maddison, 2009)). With each species tree, we generated 500 random gene trees under the coalescent process within the species tree using `Mesquite`. To generate the sequences with each gene tree, we used the software `PAML` (Yang, 1997) under the Jukes-Cantor (JC) $+\Gamma$ model (Jukes and Cantor, 1969) and the GTR $+\Gamma$ model. For the set of gene trees under the first species tree, we generated sequences under the GTR $+\Gamma$ model with rate parameters set as 10, 5, 1, 2, 3 (as the order of the input for the software `PAML` and the other parameter is dependent of these five parameters) and the number of categories for the discrete gamma model is 1 with $\alpha = 1.0$. For the other set of gene trees under the second species tree, the substitution rate for the JC model was set equal to 1 and the number of categories of the discrete gamma model is 4 with $\alpha = 0.5$. The frequencies for T, C, A, and G in the data are set as 0.15, 0.35, 0.15, 0.35, respectively. We set the length of sequences as 500. To reconstruct trees from these DNA sequences, we used the NJ algorithm with the p-distance (Saitou and Nei, 1987) (referred to as

NJp method from hereafter) to reconstruct the NJ trees, and used the software `PHYML` (Guindon and Gascuel, 2003) to reconstruct MLE trees under the GTR model (Felsenstein, 1981), the Hasegawa-Kishino-Yano (HKY) model (Hasegawa *et al.*, 1985), the Kimura 2 parameter (K80) model (Kimura, 1980), and the Jukes-Cantor (JC) model (Jukes and Cantor, 1969); they are denoted by MLE-GTR, MLE-HKY, MLE-K80, and MLE-JC, respectively, from now on.

Fig. 1 shows the rates of correctly clustered genes by our three proposed clustering schemes: direct Ncut, KPCA Ncut, and t-SNE Ncut. The accuracy is calculated as follows: first, each cluster is assigned to one of the species tree by majority of genes, and then, a gene in the cluster is considered to be correctly clustered if the gene was generated from the species tree. Generally the accuracy is higher for larger species depths (larger c), which imply clearer separation. There is a significant difference of accuracy between NJp and MLE tree reconstruction methods; the NJp method (red lines) gives better clustering performance in the Ncut and k-means methods. It is also noted that the accuracy for the MLEs has clear groups depending on the dimension reduction schemes; t-SNE Ncut (doted lines), direct Ncut (dashed lines), and KPCA Ncut (solid lines) give groups of similar accuracy levels in this order.

To show the advantage of using the BHV tree space over the Euclidean space, we applied the same clustering methods to Euclidean distance matrices $D(T)$ in $\mathbb{R}^{\binom{n}{2}}$, and compared the clustering accuracy obtained. The differences (which are easily noted) are shown in Table 1, however we focus only on depth ratios $c = 0.6$ and 1.2 for ease of presentation; the remaining depths also portray the same differences and can be found in the Supplementary Material (which is freely available at `http://polytopes.net/research/cluster/`). We observe that in most cases, the BHV tree space gives better clustering accuracy than when using Euclidean distances. Even though the Euclidean distance of MLE-GTR with t-SNE for $c = 0.6$ and t-SNE and KPCA for $c = 1.2$ gives more accurate clustering, these accuracies are much lower than the ones obtained by NJp.

To show the performance of the normalized cut framework compared to other standard clustering methods in the field, we performed the same experiments using the well-known k-means and hierarchical clustering. Note that k-means clustering is infeasible for the BHV tree space, but instead it can be applied with a dimension reduction scheme before putting it to the test. As such, there is no *direct* application in the results shown in Fig. 2.

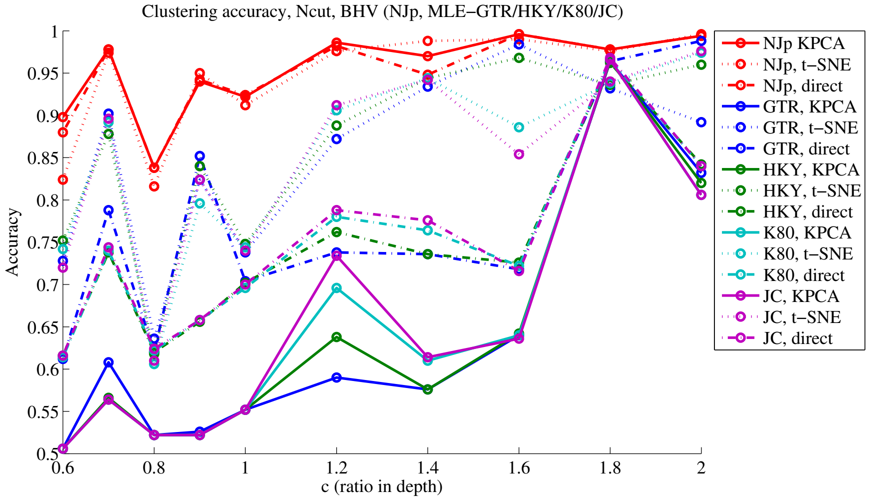

Figure 1: Ncut Clustering accuracy for simulated data. NJp gives superior accuracy than MLE. The results of MLE show three groups depending on the three clustering methods.

From Figures 1, 2, and 3, we observe that k-means is indeed a viable option for accurately clustering the trees, performing similarly to our proposed normalized cut framework. On the other hand, hierarchical clustering has proven to be worse in clustering in this context: more specifically, KPCA and direct methods result in significantly imbalanced clusters, and t-SNE shows improvement but is still worse than both Ncut and k-means. Even when other linkage methods were implemented, the results were similar. This could be caused by the agglomerative nature of the method; if two clusters have some overlap, it will easily fail. Note that Gori *et al.* (2015) also reported unfavorable results for hierarchical clustering. From our computational study, it can be concluded that normalized cut is very effective in reproducing the cluster structure in gene trees when using BHV distances; moreover, the NJp method is superior in recognizing clusters when compared to MLE methods. Hierarchical clustering, on the other hand, is not recommended in this context. Last, a main advantage of the normalized cut framework is that it requires no dimension reduction, but instead can be directly applied on the BHV space. Computational results of clustering accuracy for k-means with Euclidean distances are also presented in Table 2; the results for hierarchical

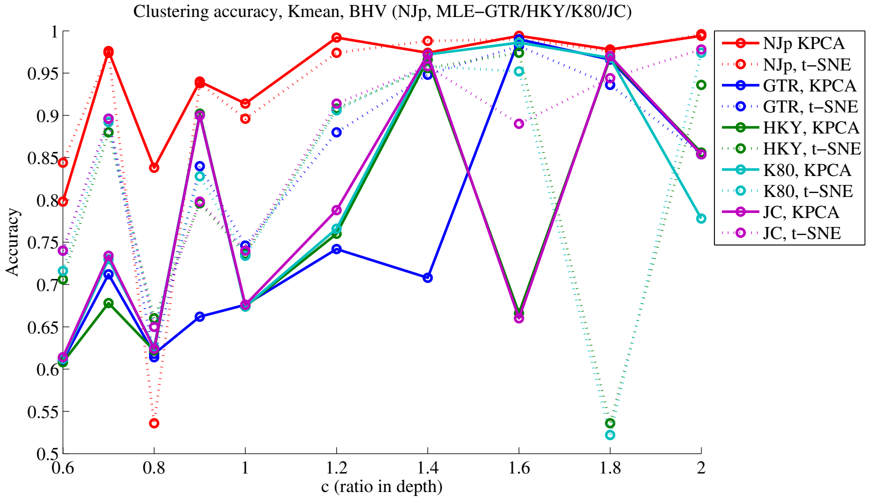

Figure 2: k-means Clustering accuracy for simulated data. It performs similarly to the normalized cut framework; the main difference is the lack of a direct application of k-means on the datasets.

clustering, as it performs poorly, are omitted here but are provided for a full computational comparison in the Supplementary Material (available at `http://polytopes.net/research/cluster/`).

In addition to the clustering accuracy, we confirmed the correctness of the species tree reconstructed from the genes in each cluster. Table 3 shows the Robinson-Foulds (RF) distances (Robinson and Foulds, 1981) between the reconstructed tree and the true species tree (T_1 and T_2) used to generate gene trees in each species depth defined above. Each cluster obtained by the normalized cut method applied directly on the set of trees reconstructed by NJp is used to reconstruct a species tree and we denote them as C_1 and C_2. We estimated each species tree by the software `PHYML` (Guindon and Gascuel, 2003) on the concatenated sequences under the GTR model. For a comparison we reconstructed the species tree from all the concatenated sequences we generated (denoted by C). To measure the accuracy, we computed the RF distance between the reconstructed species tree (C_i) and each of T_1 and T_2, as well as the RF distance between the two true species trees (D_{T_1,T_2}). We can see that D_{C_1,T_1} and D_{C_2,T_2} are significantly smaller than the RF distances D_{C,T_1} and D_{C,T_2}, which demonstrates the effectiveness of

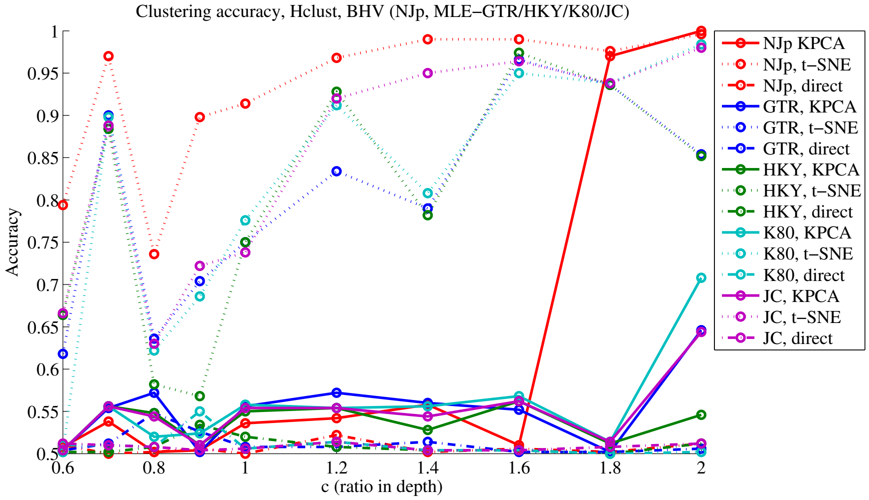

Figure 3: Hierarchical Clustering accuracy for simulated data. The results, albeit similar, are still statistically worse than the other two clustering techniques.

the clustering approach to species tree reconstruction in phylogeny.

To see how effectively the employed reconstruction methods for gene trees work, we also compare the accuracy between the true gene tree and each of the reconstructed trees under different reconstruction methods. The results, which validate that NJp performs better than, or at least as well as, the other methods, are shown in Table 4.

3.2 Genome data set on coelacanths, lungfishes, and tetrapod

On top of the simulated datasets, we have also applied the clustering methods to the dataset comprising 1,290 nuclear genes encoding 690,838 amino acid residues obtained from genome and transcriptome data by Liang *et al.* (2013). Over the last decades, the phylogenetic relations between coelacanths, lungfishes, and tetrapods have been controversial despite the existence of many studies on them (Hedges, 2009). Most morphological and paleontological studies support the hypothesis that lungfishes are closer to tetrapods than they are to coelacanths (Tree 1 in Figure 1 from Liang *et al.* (2013)), however,

	NJp			MLE-GTR		
	KPCA	t-SNE	direct	KPCA	t-SNE	direct
BHV	0.898	0.848	0.880	0.506	0.722	0.612
Euclid	0.666	0.732	0.782	0.504	0.750	0.536

(a) $c = 0.6$

	NJp			MLE-GTR		
	KPCA	t-SNE	direct	KPCA	t-SNE	direct
BHV	0.986	0.970	0.982	0.590	0.886	0.736
Euclid	0.960	0.962	0.962	0.824	0.934	0.680

(b) $c = 1.2$

Table 1: Comparison of clustering accuracy between BHV space and Euclidean space when using *normalized cut*. Euclidean distances give worse results than geodesic distances in the BHV tree space. BHV geodesic distance with NJp tree construction seems to be the most suitable for clustering.

	NJp			MLE-GTR		
	KPCA	t-SNE	direct	KPCA	t-SNE	direct
BHV	0.780	0.824	N/A	0.582	0.744	N/A
Euclid	0.618	0.738	0.782	0.504	0.718	0.548

(a) $c = 0.6$

	NJp			MLE-GTR		
	KPCA	t-SNE	direct	KPCA	t-SNE	direct
BHV	0.992	0.972	N/A	0.742	0.870	N/A
Euclid	0.966	0.964	0.966	0.856	0.934	0.682

(b) $c = 1.2$

Table 2: Comparison of clustering accuracy between BHV space and Euclidean space when using *k-means*. We observe that the normalized cut framework and k-means both perform well. There is no direct application of k-means on the BHV space, which is a main advantage of the normalized cut.

there exists research in the field that supports the hypothesis that coelacanths are closer to tetrapods (Tree 2 in Figure 1 from Liang *et al.* (2013)). Others support the hypothesis that coelacanths and lungfishes form a sister clades (Tree 3 in Figure 1 from Liang *et al.* (2013)) or tetrapods, lungfishes, and coelacanths cannot be resolved (Tree 4 in Figure 1 from Liang *et al.* (2013)). In this subsection, we apply the normalized cut framework for clustering to the genome data set from Liang *et al.* (2013) and analyze each obtained cluster.

We applied the clustering methods (with and without a dimension reduction) to the distance matrix computed from the set of gene trees constructed by the NJp method. The number of clusters in Ncut was set to two, that is, a bipartition, which is shown to be reliable, as discussed later. Fig. 4 shows the

c	D_{C,T_1}	D_{C,T_2}	D_{T_1,T_2}	D_{C_1,T_1}	D_{C_2,T_2}
0.6	10	8	14	0	4
0.7	6	12	14	0	2
0.8	10	4	14	2	0
0.9	6	8	14	6	0
1.0	10	6	14	2	2
1.2	10	6	14	2	4
1.4	6	14	14	4	2
1.6	12	6	12	4	2
1.8	4	10	12	0	0
2.0	8	12	12	2	6

Table 3: Robinson-Foulds (RF) distances between the reconstructed species tree and the true spices tree. The symbols C, C_1, C_2 denote the reconstructed species trees by concatenation of all genes, genes in cluster 1, and genes in cluster 2, respectively. T_1 and T_2 are the true species trees.

clustering results with KPCA and t-SNE, plotted on the three-dimensional space found by the dimension reduction. The red and blue colors show the two clusters, where the color density represents the bootstrap confidence explained below.

To evaluate the stability of clustering, we computed the bootstrap confidence probability for each gene. Namely, given an $N \times N$ distance matrix (D_{ij}) as input to the Ncut, we generated random resampling $\{i_1, \ldots, i_N\}$ from $\{1, \ldots, N\}$ with replacement, and applied Ncut to $(D_{i_a i_b})_{a,b=1}^{N}$. We repeated this procedure 100 times with independent random indices, and computed the ratio that a gene is classified in the same cluster as the one given by (D_{ij}).

We computed the bootstrap confidence for all 1,290 genes. The cumulative distribution functions of these values are shown for the tree clustering methods in Fig. 5 (left). The ratio of genes with confidence above 0.95 is 91.4%, 83.8%, and 99.2% for direct Ncut, KPCA Ncut, and t-SNE Ncut, respectively. For comparison, we computed the bootstrap confidence for Ncut with three clusters. Fig. 5 (right) shows the cumulative distribution function, in which the two cluster case has high bootstrap confidence, whereas using three clusters has a much lower confidence, revealing that the latter is unstable. From these observations, we see that the two clusters obtained by

c	NJp	ML-JC	ML-K80	ML-HKY	ML-GTR
0.6	1.67 (0.45)	1.80 (0.44)	1.78 (0.44)	1.79 (0.44)	1.69 (0.45)
0.7	1.41 (0.52)	1.52 (0.48)	1.52 (0.48)	1.49 (0.49)	1.47 (0.49)
0.8	1.72 (0.43)	1.77 (0.42)	1.75 (0.42)	1.72 (0.42)	1.66 (0.44)
0.9	1.67 (0.44)	1.76 (0.42)	1.75 (0.42)	1.76 (0.42)	1.71 (0.44)
1.0	1.91 (0.42)	1.91 (0.42)	1.91 (0.42)	1.94 (0.42)	1.88 (0.43)
1.2	1.38 (0.50)	1.60 (0.45)	1.60 (0.46)	1.61 (0.45)	1.60 (0.45)
1.4	1.67 (0.42)	1.68 (0.44)	1.66 (0.44)	1.64 (0.46)	1.61 (0.46)
1.6	1.35 (0.50)	1.49 (0.47)	1.47 (0.48)	1.46 (0.47)	1.42 (0.48)
1.8	1.79 (0.41)	1.70 (0.44)	1.66 (0.44)	1.71 (0.43)	1.67 (0.45)
2.0	1.28 (0.50)	1.43 (0.48)	1.40 (0.49)	1.39 (0.49)	1.30 (0.51)
Ave.	1.58 (0.46)	1.67 (0.45)	1.65 (0.45)	1.65 (0.45)	1.60 (0.46)

Table 4: Comparison between the accuracy of the reconstructed gene tree under different methods and the true gene tree. The value in each cell represents $P_c\,(d_T)$ where P_c is the average RF distance among 1,000 gene trees and d_T is the topological error index, i..e., the proportion of the correct tree topology each method reconstructed among 1,000 gene trees. This result validates that, on average, NJp outperforms the other reconstruction techniques.

the methods are not artifacts but instead a stable structure in the genome data.

The clusters obtained by the three methods look different in their shapes. We then examined agreements of the clusters at the gene level. After extracting the genes with bootstrap confidence not less than TH (TH = 0.90 or 0.95), we evaluated the agreement of methods A and B by

$$t_{AB} := \frac{|C_A^1 \cap C_B^1| + |C_A^2 \cap C_B^2|}{N_A},$$

where N_A is the number of genes by Method A with confidence larger than TH and C_A^i is the i-th ($i = 1, 2$) cluster by Method A ($N_A = |C_A^1| + |C_A^2|$). We identified which cluster in A corresponds to a cluster in B based on the number of common genes. Table 5 shows the value t_{AB} for every pair of the three methods. We can see that majority of genes in a cluster agrees to another cluster given by a different method. This confirms that the clustering reveals the structure of the data. KPCA Ncut and t-SNE Ncut are slightly less consistent, which may be caused by the difference of N_A for the two methods.

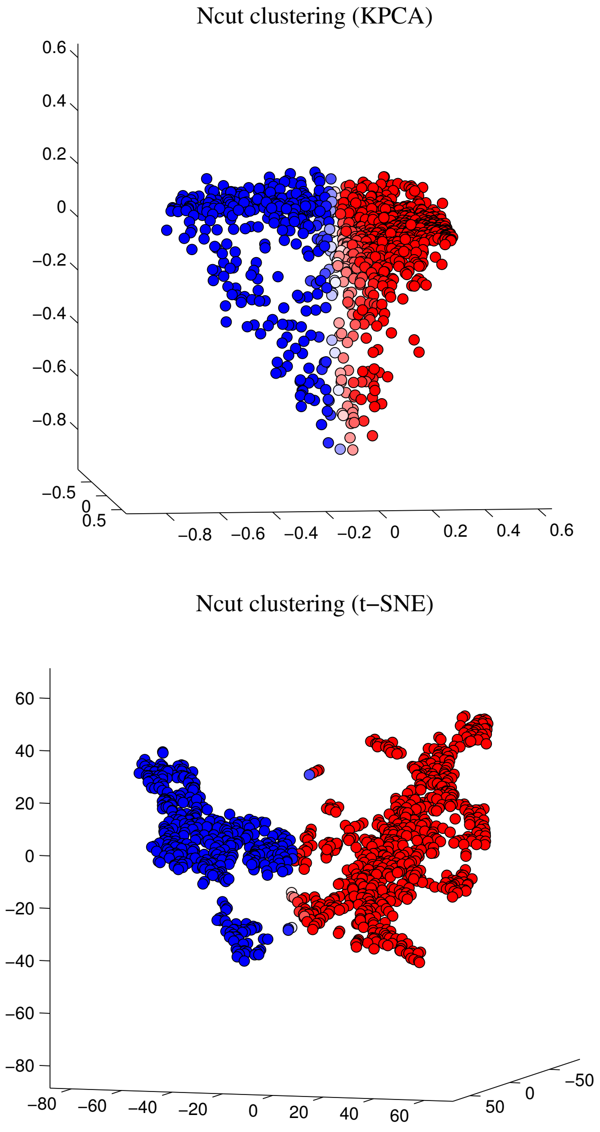

Figure 4: Clustering of the genome data set. The two clusters are depicted in red and blue with bootstrap confidence shown by color density.

Finally we conducted the phylogenetic analysis on the clusters of gene trees. For each clustering method (direct Ncut, KPCA Ncut, and t-SNE Ncut), we have reconstructed a consensus tree from each cluster. To construct

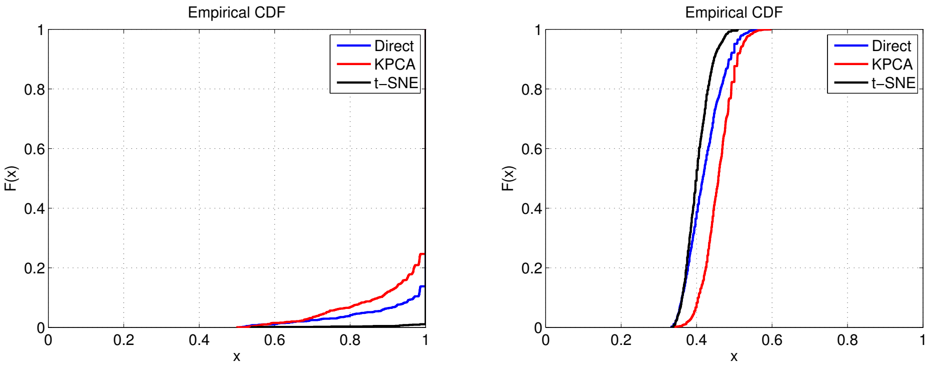

Figure 5: Cumulative distribution functions of bootstrap confidence values for clustering. The two clusters (left) are reliable, while the three clusters (right) are unstable.

the consensus tree, we have used the gene trees in each cluster with bootstrap confidence greater than 0.95 and took the majority rule with more than 50% for reconstructing the consensus tree for resolving each split on the tree. With all the clustering methods, the result suggests that there are two clusters in the genome-wide data set on coelacanths, lungfishes, and tetrapods: the number of genes are $(858, 322)$, $(761, 320)$, and $(817, 463)$ for direct Ncut, KPCA Ncut and t-SNE Ncut, respectively.

With all of the three methods, direct Ncut, Ncut with KPCA, and Ncut with t-SNE, one cluster of the gene trees provides the tree topology Tree 4 from Figure 1 in Liang *et al.* (2013), while the other cluster gives the tree topology Tree 2 from Figure 1 in Liang *et al.* (2013) (see Fig. 6).

We have also reconstructed a tree from each cluster by concatenating the alignments using the software `PhyloBayes 3.3` under a mixture model CAT $+\Gamma 4$ with two independent MCMC runs for 10,000 cycles. However, we did not observe any difference in the tree topologies, i.e., the reconstructed trees all have the same tree topology as Tree 1 from Figure 1 in Liang *et al.* (2013) (see Fig. 7).

4 Discussion

In this paper, we have shown three main computational results: first, the Ncut clustering algorithm works well on the set of gene trees reconstructed

(a) TH=0.90

$A \backslash B$	direct	KPCA	t-SNE	N_A
Direct	-	0.917	0.800	1207
KPCA	0.912	-	0.757	1135
t-SNE	0.812	0.785	-	1284

(b) TH = 0.95

$A \backslash B$	direct	KPCA	t-SNE	N_A
Direct	-	0.896	0.786	1180
KPCA	0.886	-	0.712	1081
t-SNE	0.803	0.757	-	1280

Table 5: Agreement of clusters among the three methods for the normalized cut. The rightmost column shows the number of selected genes for each method (N_A).

via the NJp under the evolutionary models; secondly, via the Ncut clustering algorithm we were able to identify two clusters on the genome data sets from Liang *et al.* (2013); last, k-means performs equally well after dimension reduction, while hierarchical clustering is always outperformed in this context. More specifically, as far as the computational experiments are concerned, we were able to show that the normalized cut framework works effectively on the set of gene trees reconstructed via the NJp method compared to the trees reconstructed via the MLE under the evolutionary models (see Table 5, Fig. 1, as well as Fig. 2 and 3 for the other methods). It is not clear why this phenomenon appears in our computational study and it is of interest to further investigate mathematically the reasons behind it.

A very important finding has to do with the dataset of *coelacanths, lungfishes, and tetrapods*. Using the Ncut algorithm on the gene trees reconstructed via the NJp method, we were able to identify two clusters. Bootstrap confidence analysis suggests that these are two reliable clusters and it appears to be very unlikely to have more than two clusters (see Fig. 5). From the two clusters we were able to find using the Ncut framework, we have reconstructed the consensus trees and their tree topologies did not support the hypothesis that lungfishes are the closest living relatives of tetrapods as in Liang *et al.* (2013), but supported the hypotheses that coelacanths are most closely related to tetrapods, and that the coelacanth, lungfish, and tetrapod lineages diverged within a very short time interval. Since clustering analysis with Ncut does not infer any evolutionary events that caused the clusters, it would be interesting and important to further investigate how these clusters were made in the evolutionary history.

Funding

This work was supported by JSPS KAKENHI 26540016. Chrysafis Vogiatzis would also like to acknowledge support from ND EPSCoR NSF #1355466.

References

Abascal, F. and Valencia, A. (2002). Clustering of proximal sequence space for the identification of protein families. *Bioinformatics*, **18**(7), 908–921.

Amemiya, C. T., Alföldi, J., and et. al. (2013). The african coelacanth genome provides insights into tetrapod evolution. *Nature*, **496**, 311–316.

Betancur, R., Li, C., Munroe, T., Ballesteros, J., and Ortí, G. (2013). Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (teleostei: Pleuronectiformes). *Systematic Biology*, page doi:10.1093/sysbio/syt039.

Billera, L., Holmes, S., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.*, **27**(4), 733–767.

Bininda-Emonds, O., Gittleman, J., and Steel, M. (2002). The (super)tree of life: procedures, problems, and prospects. *Annu. Rev. Ecol. Syst.*, **33**, 265–289.

Bollback, J. and Huelsenbeck, J. (2009). Parallel genetic evolution within and between bacteriophage species of varying degrees of divergence. *Genetics*, **181**(1), 225–234.

Brito, P. and Edwards, S. (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Carballido-Gamio, J., Belongie, S., and Majumdar, S. (2004). Normalized cuts in 3-d for spinal mri segmentation. *Medical Imaging, IEEE Transactions on*, **23**(1), 36–44.

Carling, M. and Brumfield, R. (2008). Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in passerina buntings. *Genetics*, **178**, 363–377.

Chatterji, S., Yamazaki, I., Bai, Z., and Eisen, J. A. (2008). Compostbin: A dna composition-based algorithm for binning environmental shotgun reads. In *Research in Computational Molecular Biology*, pages 17–28. Springer.

Chen, D., Burleigh, G. J., and Fernández-Baca, D. (2007). Spectral partitioning of phylogenetic data sets based on compatibility. *Systematic biology*, **56**(4), 623–632.

Cox, I. J., Rao, S. B., and Zhong, Y. (1996). "ratio regions": A technique for image segmentation. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 2, pages 557–564. IEEE.

Edwards, S. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, **63**, 1–19.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis" (5th ed)*. Wiley.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Fritzsch, B. (1987). The inner ear of the coelacanth fish latimeria has tetrapod affinities. *Nature*, **327**, 153–154.

Gori, K., Suchan, T., Alvarez, N., Goldman, N., and Dessimoz, C. (2015). Clustering genes of common evolutionary history. *arXiv preprint arXiv:1510.02356*.

Gorr, T., Kleinschmidt, T., and Fricke, H. (1991). Close tetrapod relationships of the coelacanth latimeria indicated by haemoglobin sequences. *Nature*, **351**, 394–397.

Gretton, A., Smola, A. J., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Scholkopf, B., and Logothetis, N. (2005). Kernel constrained covariance for dependence measurement. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5), 696–704.

Hartigan, J. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, **22**, 160–174.

Haws, D., Huggins, P., O'Neill, E. M., Weisrock, D. W., and Yoshida, R. (2012). A support vector machine based test for incongruence between sets of trees in tree space. *BMC Bioinformatics*, **13**(210). doi:10.1186/1471-2105-13-210.

Hedges, S. (2009). Vertebrates (vertebrata). *In: Hedges SB, Kumar S, editors. The timetree of life*, pages 309–314.

Heled, J. and Drummond, A. (2011). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**(3), 570–580.

Hess, J. and Goldman, N. (2011). Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: yeasts revisited. *PLoS ONE*, **6**, e22783.

Higham, D., Kalna, G., and Kibble, . (2007). Spectral clustering and its use in bioinformatics. *Journal of Computational and Applied Mathematics*, **204**(1), 25 – 37. Special issue dedicated to Professor Shinnosuke Oharu on the occasion of his 65th birthday.

Hochbaum, D. S. (2010). Polynomial time algorithms for ratio regions and a variant of normalized cut. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **32**(5), 889–898.

Hochbaum, D. S. (2013). A polynomial time algorithm for rayleigh ratio on discrete variables: Replacing spectral techniques for expander ratio, normalized cut, and cheeger constant. *Operations Research*, **61**(1), 184–198.

Holmes, S. (2005). Statistical approach to tests involving phylogenies. In O. Gascuel, editor, *Mathematics of Phylogeny and Evolution*, chapter 4, pages 91–117. Oxford University Press, New York.

Huson, D. H., Klopper, T., Lockhart, P. J., and Steel, M. A. (2005). *Reconstruction of reticulate networks from gene trees*. Research in Computational Molecular Biology, Proceedings. Springer-Verlag Berlin, Berlin.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends Genet.*, **22**, 225–231.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–32. New York Academic Press.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Leigh, J. W., Lapointe, F.-J., Lopez, P., and Bapteste, E. (2011). Evaluating phylogenetic congruence in the post-genomic era. *Genome biology and evolution*, **3**, 571–587.

Liang, D., Shen, X., and Zhang, P. (2013). One thousand two hundred ninety nuclear genes from a genome-wide survey support lungfishes as the sister group of tetrapods. *Mol Biol Evol.*, **30**(8), 1803–1807.

Liu, K., Raghavan, S., Nelesen, S., Linder, C., and Warnow, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, **324**, 1561–1564.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, **46**(3), 523–536.

Maddison, W. P. and Maddison, D. (2009). Mesquite: a modular system for evolutionary analysis. version 2.72. Available at `http://mesquiteproject.org`.

Maimon, O. and Rokach, L. (2005). *Data mining and knowledge discovery handbook*, volume 2. Springer.

Martin, A. P. and Burg, T. M. (2002). Perils of paralogy: Using hsp70 genes for inferring organismal phylogenies. *Systematic Biology*, **51**, 570–587.

Miller, E., Owen, M., and Provan, J. S. (2015). Averaging metric phylogenetic trees. *Advances in Applied Mathematics*, **68**, 51–91.

Newman, M. E. J. (2013). Spectral methods for community detection and graph partitioning. *Phys. Rev. E*, **88**, 042822.

Neyman, J. (1971). Molecular studies of evolution: a source of novel statistical problems. *in Statistical Decision Theory and Related Topics*, pages 1–27. edited by Shanti S. Gupta and James Yackel.

Owen, M. and Provan, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE ACM T COMPUT BI*, **8**(1), 2–13.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.*, **5**, 568–583.

Posada, D. and Crandall, K. (2002). The effect of recombination on the accuracy of phylogeny reconstruction. *Journal of Molecular Evolution*, **54**, 396–402.

Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA*, **95**(11), 6239–6244.

Robinson, D. and Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of alignments can be positively misleading. *Theoretical Population Biology*, **100**, 56–62.

Saitou, N. and Nei, M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319.

Sharon, E., Galun, M., Sharon, D., Basri, R., and Brandt, A. (2006). Hierarchy and adaptivity in segmenting visual scenes. *Nature*, **442**(7104), 810–813.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 888–905.

Takahata, N. (1989). Gene genealogy in 3 related populations: consistency probability between gene and population trees. *Genetics*, **122**, 957–966.

Takahata, N. and Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, **124**, 967–978.

Takezaki, N., Figueroa, F., Zaleska-Rutczynska, Z., Takahata, N., and Klein, J. (2004). The phylogenetic relationship of tetrapod, coelacanth, and lungfish revealed by the sequences of forty-four nuclear genes. *Mol Biol Evol*, **21**, 1512–1524.

Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lec. Math. Life Sci.*, **17**, 57–86.

Taylor, J. W., Jacobson, D. J., Kroken, S., Kasuga, T., Geiser, D. M., Hibbett, D. S., and Fisher, M. C. (2000). Phylogenetic species recognition and species concepts in fungi. *Fungal Genetics and Biology*, **31**, 21 – 32.

Thompson, K. and Kubatko, L. (2013). Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies. *BMC Bioinformatics*, **14**, 200.

van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Weisrock, D. W., Shaffer, H. B., Storz, B. L., Storz, S. R., Storz, S. R., and Voss, S. R. (2006). Multiple nuclear gene sequences identify phylogenetic species boundaries in the rapidly radiating clade of mexican ambystomatid salamanders. *Molecular Ecology*, **15**, 2489–2503.

Weyenberg, G., Huggins, P., Schardl, C., Howe, D., and Yoshida, R. (2014). kdetrees: non-parametric estimation of phylogenetic tree distributions. *Bioinformatics*, **30**(16), 2280–2287.

Xing, E. and Karp, R. (2001). CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, **17**(suppl 1), S306–S315.

Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *CABIOS*, **15**, 555–556.

Yao, W., Krzystek, P., and Heurich, M. (2012). Tree species classification and estimation of stem volume and {DBH} based on single tree extraction by exploiting airborne full-waveform lidar data. *Remote Sensing of Environment*, **123**(0), 368 – 380.

Yu, Y., Warnow, T., and Nakhleh, L. (2011). Algorithms for mdc-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *J Comput Biol*, **18**(11), 1543–1559.

Zhang, S.-B., Zhou, S.-Y., He, J.-G., and Lai, J.-H. (2011). Phylogeny inference based on spectral graph clustering. *Journal of Computational Biology*, **18**(4), 627–637.

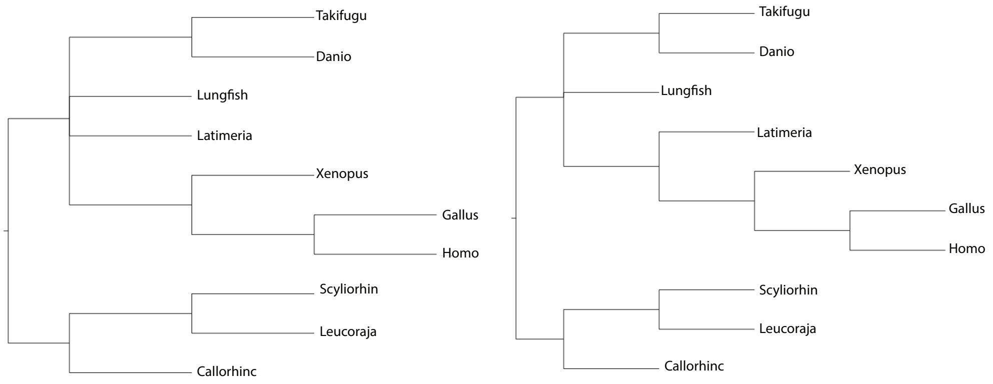

Figure 6: The majority rule consensus tree consists of gene trees with more than 0.95 bootstrap values in each cluster. Each split in the trees is resolved only if we have majority, i.e. 50% of all given gene trees in each set agree. All the three clustering methods give the same two spices trees.

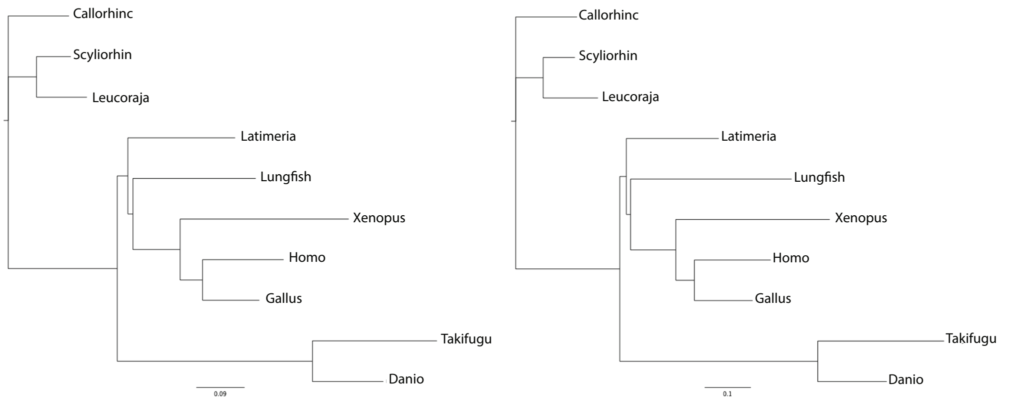

Figure 7: The reconstructed trees obtained by concatenating the alignments from each cluster after using direct Ncut. For this result, we employed the Bayesian inference using the software `PhyloBayes 3.3` under a mixture model CAT $+\Gamma 4$ with two independent MCMC runs for 10,000 cycles.