2018-10-08

Multi loci phylogenetic analysis with gene tree clustering

Yoshida, Ruriko; Fukumizu, Kenji; Vogiatzis, Chrysafis

Multi Loci Phylogenetic Analysis with Gene Tree Clustering

Kenji Fukumizu   Ruriko Yoshida   Chrysafis Vogiatzis

October 8, 2018

Abstract

Summary: Both theory and empirical evidence indicate that phylogenies (trees) of different genes (loci) do not display precisely matched topologies. This phylogenetic incongruence is attributed to the reticulated evolutionary history of most species due to meiotic sexual recombination in eukaryotes, or horizontal transfers of genetic materials in prokaryotes. Nonetheless, most genes do display topologically related phylogenies; this implies they form cohesive subsets (clusters). In this work, we compare popular clustering methods, and show how the performance of the normalized cut framework is 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. We proceed to present 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.

Results: First, we show using simulated data that indeed the Ncut framework accurately clusters the set of gene trees given a species tree under the coalescent process. We then 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 stellar performance of the Ncut framework even without dimension reduction, the similar performance portrayed by k-means and Ncut under most dimension reduction schemes, the utter failure of hierarchical clustering to accurately capture clusters, as well as the significantly better performance of the NJp method, as compared to MLE.

1　Introduction

During especially this last decade, the field of phylogenetics has been undergoing a gradual paradigm 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 (Huson *et al.*, 2005; Weisrock *et al.*, 2006; Taylor *et al.*, 2000). 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 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 multi-locus approaches (e.g. Carling and Brumfield (2008); Yu *et al.* (2011); Betancur *et al.* (2013); Heled and Drummond (2011); Thompson and Kubatko (2013)).

There exist numerous processes that can reduce the correlation among gene trees. 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). 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 (Posada and Crandall, 2002; Martin and Burg, 2002; Edwards, 2009). 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. These examples suggest that the distribution of eukaryotic 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, we first need to find how many components of distributions there are 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".

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, 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) etc, the MLE on a sequence data concatenated across genes under the assumption that all sites have evolved independently and identically on a fixed tree is a statistically inconsistent estimator of the species tree.

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*. 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 found in the matrix. 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.

This paper focuses on such tree spaces and presents a computational study on multi-loci phylogenetic analysis using gene tree clustering over the BHV tree space. More specifically, this work presents the differences and the performance of a normalized cut framework, k-means, and hierarchical clustering in the Euclidean and the BHV tree space, and under different dimension reduction approaches using simulated datasets. The paper is organized as follows. Section 2 offers a basic review of the BHV space, the normalized cut framework, and the dimension reduction for the interested reader. In Section 3, we present our computational study and the results we obtained using simulated datasets. Moreover, Section 4 discusses the results and focuses on our main computational results, while also summarizing our work and offering insight for possible future directions.

2   Fundamentals

In this paper, we present a comparative study of different methods for multi-loci phylogenetic analysis using gene tree clustering based on the distance matrix obtained by the geodesic distances between two trees 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., Schenker *et al.* (2003)), and hierarchical clustering (the interested reader is 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 quality of solutions compared to applying our clustering methods directly upon the original distance matrix. This reduction also helps with the visualization of the data. 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. Hereafter, we refer to the above three branches as *direct, KPCA*, and *t-SNE*, respectively.

We now proceed to offer some basics on the BHV tree space, the nor-

malized cut problem, and the different dimension reduction techniques used herein.

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. (Unrooted trees can be accommodated by using either the Ferras transform, or by designating an arbitrary 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 non-negative.) For each topology, the orthant coordinates correspond to edge lengths in the tree.

Since all tree topologies have the same set of n terminal leaves, and each of these leaves is associated with a single terminal edge, the orthant coordinates associated with the terminal edges are of less interest than those of internal nodes. As a result, we will simplify our discussion by ignoring the terminal edge lengths, and concern ourselves primarily with the portion of each orthant which describes the internal edges. (Recall 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.

Since each orthant is locally 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, we could do this through 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 matter.

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. This is also helpful for visualization purposes. For the details of the clustering and dimension reduction methods, see the Supplementary Material.

There are many standard clustering methods raging from non-hierarchical clustering such as k-means to hierarchical clustering methods. Among others, this paper considers three methods: normalized cut, k-means, and hierarchical clustering (average linkage). The k-means is the most standard non-hierarchical clustering method, which has been used in a large number of applications. Note, however, that with BHV geodesic distance the k-mean methods is not computationally feasible, because update of centroids required in the methods is too expensive. As a linkage method for hierarchical clustering we use average linkage, since it is applied to general distance or dissimilarity measures.

From the clustering methods presented in this computational study, the normalized cut framework (Shi and Malik, 2000) is applied for the first time in phylogenetics, to the best of our knowledge. Normalized cut can be employed for clustering using only a similarity or dissimilarity matrix; we can observe that 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). 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 (1) is minimized with $S \cup \bar{S} = V$ and $S \cap \bar{S} = \emptyset$.

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

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. The normalized cut framework 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).

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. In our experiments, for simplicity, the similarity w_{ij} is given by $w_{ij} = \exp(-\frac{1}{2\sigma^2}D_{ij}^2)$, where D_{ij} is the distance matrix, and $\sigma = 1.2 \times \text{Median}\{D_{ij} \mid i \neq j\}$.

2.3   Dimension reduction

As an optional procedure 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 applied also spectral methods and Isomap, but the results were less favorable than KPCA and t-SNE.) Those methods extracted three dimensional expression of data, when applied, and the Ncut was 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 nonlinear mapping of the original data to a feature space. The nonlinear map is defined by a *positive definite kernel*, and the feature space is a possibly infinite dimensional Hilbert space provided implicitly by the positive definite kernel. KPCA gives nonlinear functions $f_1, \dots f_d$ of data points $(X_i)_{i=1}^N$ so that $(f_1(X_i), \dots, f_d(X_i))_{i=1}^N$ 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].

---

[1]While this kernel with an arbitrary distance matrix D is not necessarily positive

7

t-SNE is a method for low-dimensional expression or visualization of high-dimensional data; it typically extracts 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 probability implies 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 a probability q_{ij} defined similarly for a pair (Y_i, Y_j) is close to (p_{ij}). The points $(Y_i)_{i=1}^N$ are found with numerical optimization to minimize the Kullback-Leibler divergence between (p_{ij}) and (q_{ij}). In our experiments, a Matlab implementation by van der Maaten (lvdmaaten.github.io/tsne/) is used. The perplexity parameter, which gives a way of determining local bandwidth parameters, is set 30 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); Liang *et al.* (2013), where it was observed that there were 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 (typically used) Euclidean space and the (better suited) BHV tree space and compared the accuracy of the two spaces for the goal of recognizing more statistically accurate clusters. Last, we compared dif-

definite, in our analysis the Gram matrices $k(X_i, X_j)$ made by the given data were positive definite.

ferent 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); 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 have fixed the population size $N_e = 10,000$ and we set the species depth $c \cdot N_e$ where $c = 0.6, 0.7, 0.8, 0.9, 1, 1.2, 1.4, 1.6, 1.8, 2$. Then for each species depth $c \cdot N_e$, we generated 100 species trees from the Yule process and we picked randomly two trees from them. With each species tree, we generated 1000 random gene trees under the coalescent process within the species tree using the software `Mesquite` (Maddison and Maddison, 2009). To generate the sequences we have used the software `PAML` (Yang, 1997) under the Jukes-Cantor (JC) (Jukes and Cantor, 1969) $+\Gamma$ model with $\kappa = 4.0$, and the number of categories of the discrete gamma model is 1 with $\alpha = 1.0$ for one set of gene trees under the first species tree and we generated sequences under the GTR $+\Gamma$ model with $\kappa = 4.0$, and the number of categories of the discrete gamma model is 1 with $\alpha = 1.0$ for the other set of gene trees under the second species tree. 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) (we call NJp method from here on) 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) (Hasegawa *et al.*, 1985) model,
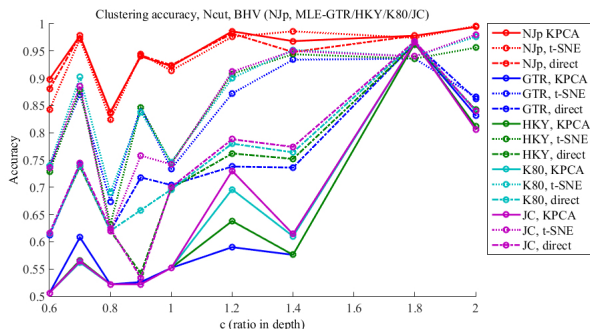
Figure 1: Ncut Clustering accuracy for simulated data. NJp gives superior accuracy than MLE. The results of MLE show three groups depending on the three clustering methods.

and the Kimura 2 parameter (K80) model (Kimura, 1980); they are denoted by MLE-GTR, MLE-HKY, and MLE-K80, respectively, from now on.

Fig. 1 shows the rates of correctly clustered genes by the three newly proposed clustering schemes: direct Ncut, KPCA Ncut, and t-SNE Ncut. Here and in the sequel, *direct* means direct application of a clustering methods without dimension reduction. Generally the accuracy is higher for larger species depths, which imply clearer separation. There is a significant difference of accuracy between NJp and MLE tree reconstruction methods; the NJp method (solid lines) gives better clustering for all the three clustering methods. It is also noted that the accuracy for the MLEs has clear groups based on the clustering schemes; t-SNE Ncut (broken lines), direct Ncut (dashed lines), and KPCA Ncut (dotted 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 $c = 0.8$ and 1.2 for ease of presentation; the remaining depths also portray the same differences and can be found in the Supplementary Material. 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-HKY with KPCA and t-SNE for $c = 0.8$ and direct for $c = 1.2$ gives more accurate clustering, these results are much lower than the ones obtained by NJp.
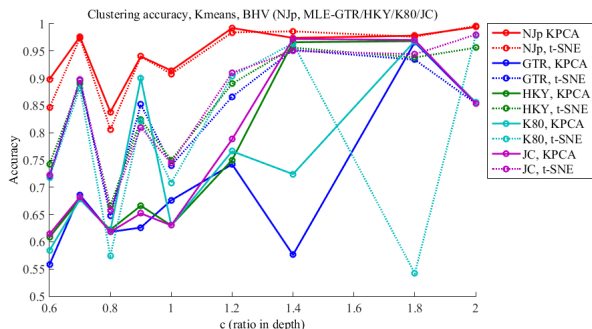
Figure 2: *k*-means Clustering accuracy for simulated data. It performs similarly to the normalized cut framework; the main difference is the lack of a direct application of *k*-means on the datasets.

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

From the Figures, we can see 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 very bad in clustering in this context, resulting in significantly imbalanced clusters. We tried other linkage methods, but the results were similar. 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 are also presented in Table 2; the results for hierarchical clustering, as it performs poorly, are omitted but are provided for completeness in the Supplementary Material.
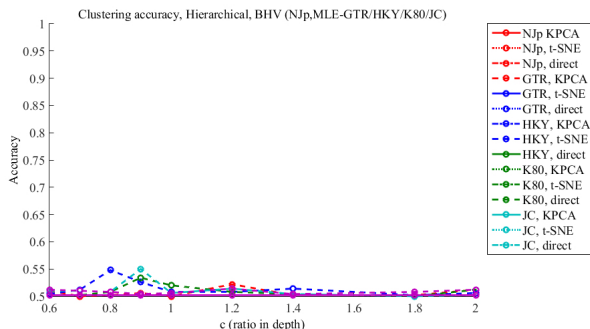
11

Figure 3: Hierarchical Clustering accuracy for simulated data. It is easy to see that the results obtained are statistically far worse than the other two main clustering techniques.

	NJp			MLE-GTR			
	KPCA	t-SNE	direct	KPCA	t-SNE	direct	(a) $c = 0.8$
BHV	0.838	0.824	0.836	0.522	0.674	0.620	
Euclid	0.840	0.860	0.750	0.740	0.826	0.740	
	NJp			MLE-GTR			
	KPCA	t-SNE	direct	KPCA	t-SNE	direct	(b) $c = 1.2$
BHV	0.988	0.976	0.982	0.872	0.944	0.738	
Euclid	0.958	0.962	0.954	0.828	0.936	0.828	

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

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 the lungfishes are closer to the tetrapods than coelacanths (Tree 1 in Figure 1 from Liang *et al.* (2013)), however, there exists research in the field that supports the hypothesis that the coelacanths are closer to the tetrapods (Tree 2 in Figure

12

	NJp			MLE-GTR			
	KPCA	t-SNE	direct	KPCA	t-SNE	direct	(a) $c = 0.8$
BHV	0.838	0.806	N/A	0.618	0.648	N/A	
Euclid	0.842	0.840	0.830	0.740	0.826	0.740	
	NJp			MLE-GTR			
	KPCA	t-SNE	direct	KPCA	t-SNE	direct	(b) $c = 1.2$
BHV	0.992	0.984	N/A	0.742	0.866	N/A	
Euclid	0.956	0.962	0.956	0.828	0.936	0.828	

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.

1 from Liang *et al.* (2013)). Others support the hypothesis that the coelacanths and the lungfishes form a sister clades (Tree 3 in Figure 1 from Liang *et al.* (2013)) or tetrapodes, 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. Fig. 4 shows the 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 a bootstrap confidence probability for each gene. Namely, given $N \times N$ distance matrix (D_{ij}) as an input to the Ncut framework, 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
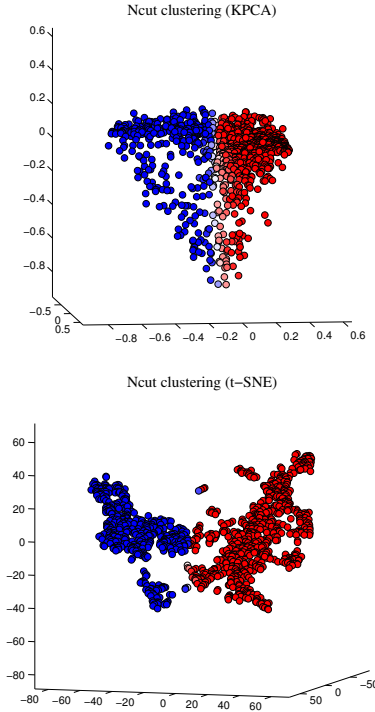
Figure 4: Clustering of the genome data set. The two clusters are depicted in red and blue with bootstrap confidence shown by color density.

function, which clearly reveals that three clusters are unstable. From these observations, we see that the two clusters obtained by the methods are not artifacts but 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 B by the number of common genes. Table 3 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
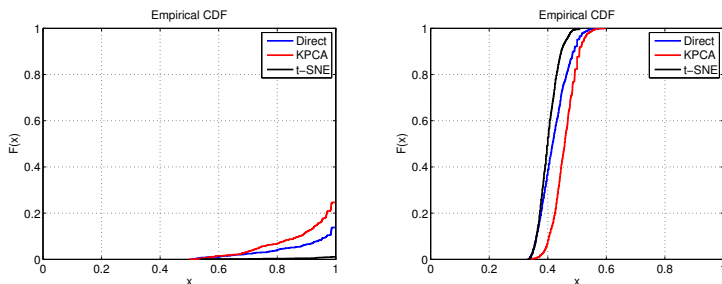
14

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

(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 3: 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).

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.

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 the consensus tree, we have used the gene trees in each cluster with bootstrap value 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. Note that we used only gene trees with bootstrap confidence ≥ 0.95.

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 $+\Gamma4$ 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 have all the same tree topology as Tree 1 in from Figure 1 in Liang *et al.* (2013) (see Fig. 7).

4   Discussion

In this paper we have shown three main results: the Ncut clustering algorithm works well on the gene trees reconstructed via the NJp under the evolutionary models; via the Ncut clustering algorithm we found 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.

Simulations: As we have shown by simulations in Section 3, 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 (Table 3, 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 investigate mathematically the reason it happens.

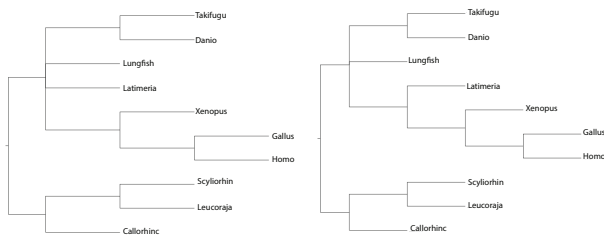*Coelacanths, lungfishes, and tetrapods data set*: Using the Ncut algorithm on

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.

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 the lungfishes are the closest living relatives of the tetrapods as in Liang *et al.* (2013), but supported the hypotheses that the coelacanths are most closely related to the 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

References

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

Schenker, A., Last, M., Bunke, H., and Kandel, A. (2003). Classification of web documents using a graph model. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 240–244. IEEE.

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.

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.
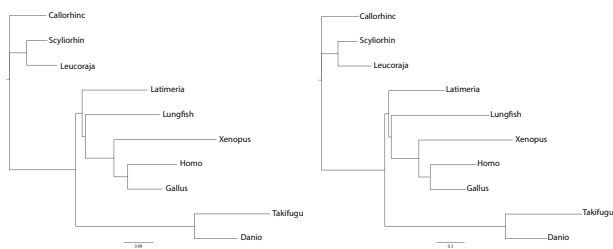
Figure 7: The reconstructed trees obtained by concatenating the alignments from each cluster after using direct Ncut. For this result, we employed the Bayesian inference using the software `PhyloBayes 3.3` under a mixture model CAT $+\Gamma4$ with two independent MCMC runs for 10,000 cycles.