Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Phylogenetic tree selection by the adjusted k-means approach

Hsiuying Wang [a] & Shan-Lin Hung [a]
[a] Institute of Statistics, National Chiao Tung University , Hsinchu , Taiwan
Published online: 31 Aug 2011.

To cite this article: Hsiuying Wang & Shan-Lin Hung (2012) Phylogenetic tree selection by the adjusted k-means approach, Journal of Applied Statistics, 39:3, 643-655, DOI: 10.1080/02664763.2011.610442

To link to this article: http://dx.doi.org/10.1080/02664763.2011.610442

Journal of Applied Statistics
Vol. 39, No. 3, March 2012, 643–655

Taylor & Francis
Taylor & Francis Group

Phylogenetic tree selection by the adjusted *k*-means approach

Hsiuying Wang* and Shan-Lin Hung

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

(*Received 3 December 2010; final version received 1 August 2011*)

The reconstruction of phylogenetic trees is one of the most important and interesting problems of the evolutionary study. There are many methods proposed in the literature for constructing phylogenetic trees. Each approach is based on different criteria and evolutionary models. However, the topologies of trees constructed from different methods may be quite different. The topological errors may be due to unsuitable criterions or evolutionary models. Since there are many tree construction approaches, we are interested in selecting a better tree to fit the true model. In this study, we propose an adjusted *k*-means approach and a misclassification error score criterion to solve the problem. The simulation study shows this method can select better trees among the potential candidates, which can provide a useful way in phylogenetic tree selection.

Keywords: phylogenetic tree; misclassification error; *k*-means; adjusted *k*-means

1. Introduction

The reconstruction of phylogenetic trees is one of the most important and interesting problems of evolutionary study [4]. There are many existing methods for constructing phylogenetic trees from molecular data: unweighted pair-group method using arithmetic averages (UPGMA), neighbor-joining, minimum evolution and maximum parsimony methods [4,11,17]. In addition, the maximum likelihood approach, Markov chain Monte Carlo-based Bayesian inference and other approaches are proposed for tree constructions [26]. Beside phylogenetic applications, the classification tree methods are useful in other applications [2,14,18–20,25].

The patterns of phylogenetic trees built from various methods may be substantially different. Selecting better trees to fit real data is an essential problem in the investigation of evolutionary processes. There are two types of errors for phylogenetic trees: the topological error and the branch-length error [22]. The former error is the differences in the branching pattern between an inferred tree and the true tree, and the latter are deviations of estimated branch lengths from the true branch lengths. Topological errors are more serious than branch-length errors. Thus, we mainly focus on topological error in the study.

*Corresponding author. Email: wang@stat.nctu.edu.tw

ISSN 0266-4763 print/ISSN 1360-0532 online

http://dx.doi.org/10.1080/02664763.2011.610442
http://www.tandfonline.com

To examine the reliability of a tree, although bootstrap methods can be used to test the reliability of a tree, they are unsuitable for tree selection. In this study, we focus on the comparison of different trees and propose a statistical method, the adjusted k-means approach, to examine the accuracy of the topology of trees and use it as a criterion to select a better tree. The trees considered in this study are based on maximum likelihood and Bayesian inference, UPGMA, neighbor-joining, minimum evolution and the maximum parsimony approaches which can be obtained by molecular evolutionary genetics analysis (MEGA) 4.1 software [12,21].

The adjusted k-means method is established by clustering multiple sequences into several groups. We then assign the sequences with the same index number in the same group and calculate the misclassification error scores for the trees. We recommend the tree with a smaller misclassification error. This approach is basically associated with a statistical method to guide the selection of better trees. The inferences for the existing approaches are usually based on an evolutionary model. From the statistical viewpoint, we can classify the sequences into several clusters by exploring the features of the sequences without assuming any evolutionary model. Since the existing estimated evolutionary models may not be close to the true evolutionary model [24], tree construction based on an evolutionary model may lead to an unsatisfactory result if the assumption of the evolutionary model is inappropriate. The method proposed in this study which does not rely on an evolutionary model is robust in selecting trees.

2. Real data example

We use the avian family Aegothelidae discussed in [3] (commonly known as owlet-nightjars) to illustrate the aim of this study. Owlet-nightjars are small nocturnal birds related to nightjars and frogmouths. Most are native to New Guinea, but some species extend to Australia, Moluccas, and New Caledonia. There is a single monotypic family Aegothelidae with the genus, *Aegotheles*. The Aegothelidae family comprises only nine extant species, all in a single genus, *Aegotheles*.

Based on mitochondrial DNA sequence, Dumbacher *et al*. [3] constructed a phylogeny of the owlet-nightjars. They analyzed mtDNA sequences cytochrome b and ATPase subunit 8 and suggested that there are nine living species of owlet-nightjars, plus one that went extinct early in the second millennium AD. They performed the maximum likelihood analyses, using likelihood heuristic searches with a 2-rate class (transitions and transversions) model of sequence evolution with gamma correction, which is identical to the HKY85 model evolution [5,6] with the addition of a gamma rate parameter [26]. The taxon used by Dumbacher *et al*. [3] includes albertisi albertisi, wallacii wallacii, wallacii gigas, etc., as shown in Table 1. The Genbank numbers for the sequences are AY090664-AY090698 (for cytochrome b) and AY090699-AY090736 (for ATPase 8). A simple form of the tree based on the results in [3] is available in tree of life web project website at http://tolweb.org/tree/, and is shown in Figure 1.

Table 1. The cluster results of the nine sequences for different k_0.

k_0	2	3	4	5	6	7
A. wallacii	1	3	2	5	4	5
A. archboldi	1	3	2	5	3	4
A. albertisi	1	3	2	5	3	4
A. albertisi salvadori	1	3	2	1	5	6
A. bennettii	1	1	1	3	2	7
A. cristatus	1	1	1	3	2	7
A. crinifrons	1	3	3	2	1	3
A. tatei	2	2	4	4	6	2
A. insignis	2	2	4	4	6	1

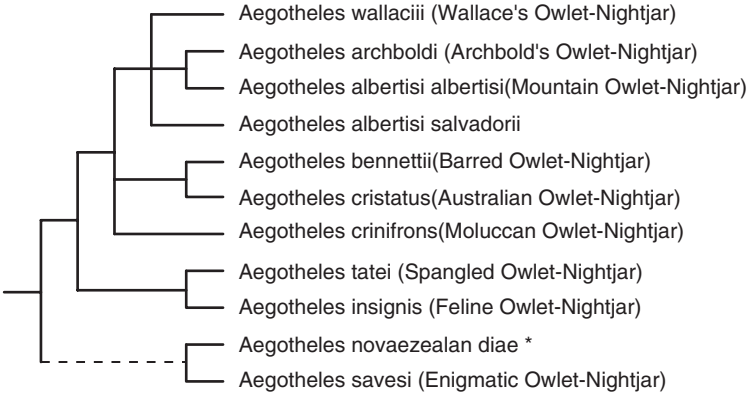

Figure 1. A simple form of the tree for the avian family Aegothelidae constructed by Dumbacher *et al*. [3]
* extinct species.

Aegotheles novaezealandiae in Figure 1 is the extinct one. According to Dumbacher *et al*. [3] the *Aegotheles savesi* may also be extinct. In this study, we exclude the two extinct species *A. novaezealandiae* and *A. savesi*, and consider the phylogenetic trees for the other nine species.

Dumbacher's tree is based on the maximum likelihood and Bayesian inference. The four phylogenetic trees based on UPGMA, neighbor-joining, minimum evolution and maximum parsimony methods for cytochrome b plotted by MEGA software are shown in Figure 2. A study related to the substitution number estimation for the example in [3] is given in [23].

Note that only UPGMA is a rooted tree. The others are unrooted trees. In a rooted tree there exists a particular node, called the root, from which a unique path leads to any other node. An unrooted tree may not have a root node. In this case, we place the root of the neighbor-joining,

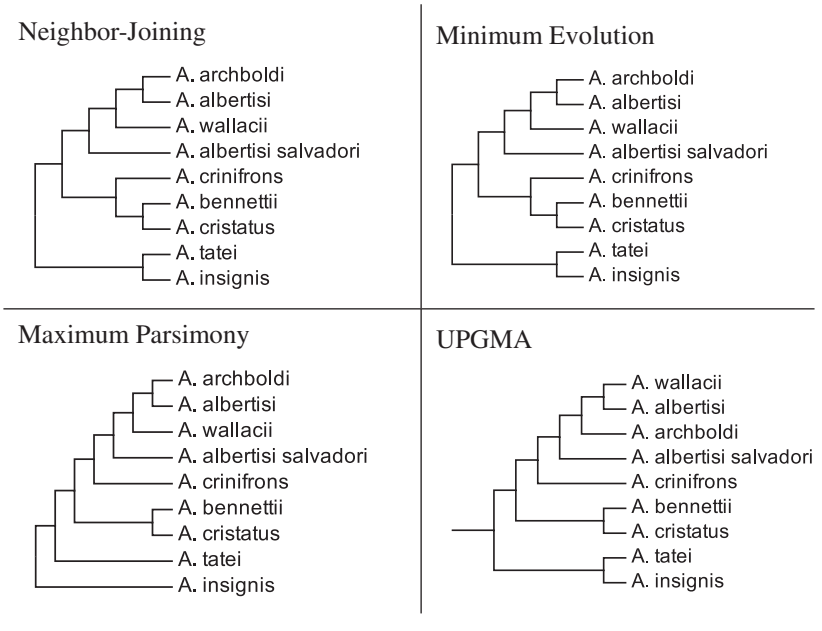

Figure 2. The neighbor-joining, minimum evolution, maximum parsimony and UPGMA trees for the avian family plotted by MEGA using nine living species.

minimum evolution and maximum parsimony trees of Figure 2 in the left-hand side as the UPGMA tree and viewed the three trees as rooted trees. In fact, the UPGMA method is a hierarchical cluster analysis which is the simplest method for tree reconstruction.

From Figure 2, it is evident that the topologies of these trees are not exactly the same. The aim of this study is to establish a reliable criterion from a statistical perspective to distinguish the inappropriate trees and select better trees. The proposed procedure for this purpose, the adjusted k-means approach, is introduced in the next section.

3. Adjusted k-means approach for categorical data

The clustering method is a useful approach for classifying the data [8–10]. The k-means clustering method proposed by MacQueen [15] and [1] is a popular clustering method to partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. However, the conventional k-means algorithm only works on numerical data, i.e. the variables are measured on a ratio scale [9]. This prohibits it from being used in applications where categorical data are involved. The nucleotide bases of a DNA sequence are A, T, C and G, which are categorical data as well as the bases in a protein sequence. The conventional k-means approach cannot be directly used to cluster the sequences.

Huang [7] proposed an extension of the k-means algorithm, the k-modes algorithm, to categorical domains. However, the k-modes algorithm may not converge. Ng *et al*. [16] provide a modified k-modes algorithm to overcome the convergence problem of the original k-modes algorithm. Although the modified algorithm may be more stable than the original k-modes algorithm, according to our computing results, it still cannot converge when it is applied to cluster the multiple nucleotide or protein sequences. Therefore, in this study, we propose an adjusted k-means algorithm to cluster the multiple sequences. The algorithm is introduced later as Procedure 1.

Before describing the approach, we first introduce some notations. First, we define the dissimilarity measure between a nucleotide or a protein sequence X and a cluster of n sequences G^n, where $X = \{x_1 x_2 \dots x_m\}$ is a nucleotide or protein sequence with length m, and $G^n = (G_1^n, G_2^n, \dots, G_n^n)$ is a set of n nucleotide or protein sequences with $G_i^n = \{g_{i1} g_{i2} \dots g_{im}\}$. Note that x_j represents the nucleotide in the jth site of the sequence X and g_{ij} represents the nucleotide in the jth site of the ith gene sequence, G_i^n, $1 \le i \le n$, $1 \le j \le m$. Then the dissimilarity measure between gene sequence X and cluster G^n is defined as follows:

$$d(X, G^n) = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} \phi(x_j, g_{ij})}{nm}, \tag{1}$$

where $\phi(x_j, g_{ij}) = I(g_{ij} \neq x_j)$ and $I(\cdot)$ denotes the indicator function.

Suppose that $G^{n_1}, G^{n_2}, \dots, G^{n_k}$ are k sets and G^{n_l} has n_l sequences, $l = 1, \dots, k$. We define the within-group measure (WGM) and the between-group measure (BGM) for $\{G^{n_1}, G^{n_2}, \dots, G^{n_k}\}$ as follows:

$$\text{WGM for } G^{n_l} \equiv \sum_{i=1}^{n_l} d\left(G_i^{n_l}, G^{n_l} \backslash G_i^{n_l}\right) \quad \text{for } 1 \le l \le k, \tag{2}$$

$$\text{BGM for } \{G^{n_{l_1}}, G^{n_{l_2}}\} \equiv \sum_{i=1}^{n_{l_1}} d(G_i^{n_{l_1}}, G^{n_{l_2}}) + \sum_{i=1}^{n_{l_2}} d(G_i^{n_{l_2}}, G^{n_{l_1}}) \quad l_1 \neq l_2 \quad \text{and} \quad 1 \le l_1, l_2 \le k, \tag{3}$$

where $A \backslash b$ denotes that set A excludes the element b.

Note that in the calculation of (1), after aligning the sequences [13], some missing sites may exist for some sequences in G^n. For a specified site, it can be classified into three cases. The first one is that a sequence, say X is missing at this site. The second case is that all sequences in G^n are missing at this site. And the third one is that some of the sequences in G^n are missing at this site, but not all sequences. In the first or the second case, the function $I(\cdot)$ in (1) for this site is defined as 0.

For the third case, we exclude the sequences with missing value at this site. Assume that the number of the sequences left is r. Then we view the group G^n as a new set G^r with r sequences and calculate the dissimilarity measure at this site. In this case, formula (1) needs to be modified to

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}\frac{\phi(x_j, g_{ij})}{m_i}\right),$$

where m_i denotes the number of sequences left excluding the sequences with the ith site missing.

3.1 *Algorithm*

Since our goal is to partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria, we prefer that the BGM is large and the WGM is small. In this case, we set up a criterion to select the clusters such that $M = \text{BGM}^* - \text{WGM}^*$ is maximum, where BGM^* denotes all BGMs of each two clusters and WGM^* denotes all WGMs described as Equations (4) and (5), that is,

$$\text{WGM for } \{G^{n_1}, G^{n_2}, \ldots, G^{n_k}\} \equiv \sum_{l=1}^{k}\sum_{i=1}^{n_l} d\left(G_i^{n_l}, G^{n_l}\backslash G_i^{n_l}\right), \tag{4}$$

$$\text{BGM for } \{G^{n_1}, G^{n_2}, \ldots, G^{n_k}\} \equiv \sum_{l_1=1}^{k}\sum_{l_2\neq l_1}^{k}\sum_{i=1}^{n_{l_1}} d(G_i^{n_{l_1}}, G^{n_{l_2}}). \tag{5}$$

In this study, based on the criterion of selecting clusters with the largest Mvalue, we provide a calculation procedure as follows.

PROCEDURE 1

Step 1. Align the n sequences, $X_1, \ldots\ldots, X_n$, and select a k value.
Step 2. Allocate every sequence to k clusters randomly.
Step 3. Allocate a sequence to the cluster with the smallest dissimilarity measure.
Step 4. Repeat Step 3 until no sequence has changed cluster after a full cycle test of the whole data set.
Step 5. Repeat Steps 2–4 s times and find the clusters in the s times with the largest M value, which are the required clusters.

Note that the results of Steps 2–4 may depend on the initial cluster selected in Step 2. The algorithm may not converge to the true clusters with maximum M value. Thus, Step 5 is provided to select the best cluster in the s trials. In our example with nine species, we have tried different s values and found that s can select around 10. The result for s being 15 or 20 is almost the same as that for s being 10 in our simulation studies. Since there are only nine species, the s value in Step 5 need not be very large. When the number of species increases, to obtain a more accurate result, we need to select a larger s value. If s is not selected to be large for large species number, the

initial cluster may seriously affect the result in Step 5. In this case, we suggest an initial partition clustered according to phylogenetic trees.

A data example is provided in the Appendix to illustrate WGM and BGM calculations.

4. Misclassification error score

In this section, we propose a method to calculate the misclassification error. Since the goal of the approach is to evaluate the performances of different tree construction methods, we can evaluate the trees under the different cluster number k and then select the tree such that it has better performance in most situations. After applying the adjusted k-means approach to cluster the n sequences to k_0 clusters, we define a misclassification error score to evaluate a tree.

We take the tree constructed by Dumbacher *et al*. [3] as an example to describe the misclassification error score calculation. Table 1 lists the cluster results by the adjusted k-means approach for $k_0 = 2, \ldots, 7$. For example, the second column in Table 1 is the clustering result of $k_0 = 3$. In this case, the sequences (*Aegotheles wallacii*, *Aegotheles archboldi*, *Aegotheles albertisi*, *Aegotheles albertisi salvadori*, *Aegotheles crinifrons*) corresponding to index 3 are clustered together in the third group; the sequences (*Aegotheles tatei*, *Aegotheles insignis*) corresponding to index 2 are clustered to the second group and so on.

We use Figure 3 to illustrate the misclassification error score calculation. The left panel of Figure 3 shows the score calculation for $k_0 = 5$. First, the sequence names in Figure 1 are replaced by the corresponding group indexes for $k_0 = 5$ in Table 1.

Define the misclassification error score as the sum of the score at each node. The score at each node is the absolute value of the difference of the group index numbers of its branches. For example, in the left panel of Figure 3, there are five nodes, A, B, C, D and E, for which we need to count the scores. Note that we do not count the score at node F because all branches are spread out from this node. Thus, it is not necessary to require a score at this node. The score value at each node is the absolute value of the difference of the index numbers from the branches spread from this node. Thus, the misclassification error score is the sum of the five scores at these five nodes, which is calculated by $0 + 0 + 0 + 8 + 1 = 9$.

The score calculation algorithm for different situations is described as follows. Note that for a node with only two branches, the score at this node is the absolute value of the difference of the

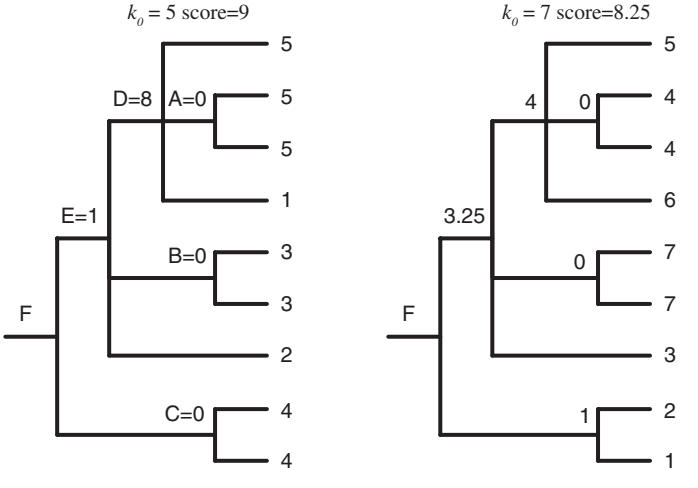

Figure 3. The tree proposed by Dumbacher *et al*. [3], excluding the two extinct species, with the index numbers replacing taxa for $k_0 = 5$ and $k_0 = 7$.

two assigned group indexes, such that node *B* has a score value of 0 with two branches whose assigned group indexes are 3 and 3. When a node has more than two branches and includes another node, such that node *D* has four branches and includes one node *A*, we need to calculate the score value at the node *A* first, and then view the two branches spread from node *A* as one single branch with assigned number 5 because the two branches both have assigned index 5. If the two branches do not have the same index, then we take the average of the indexes to be the assigned number. Thus, after viewing the two branches spread from node *A* as a single branch, node *D* has three branches, say branches 1–3. And we calculate the sum of the absolute values of the difference of the indexes number for any two of three branches indexes as this node's score (see Figure 4).

Finally, when a node has included more than one node and one single branch such as node *E*, we first calculate the average number of the branches of node *D* and node *B*, and then take the absolute value of the difference of the average number $((5+1)/2+(3+3)/2)/2$ and the index number 2 of the signal branch as the score at node *E*. Here, the average number $((5+1)/2+(3+3)/2)/2$ views the two branches at node *B* as a single branch and the two branches (taxon 1 and taxon 4 in Figure 4) as a single branch. We do not need to consider taxons 2 and 3 here because their score has been considered at node *A*.

Besides the above proposed scoring method, other scoring methods can be adopted. In this study, we do not focus on comparing different scoring methods; and exploring an optimal scoring algorithm is an interesting future research topic.

With the definition of the misclassification error score, we can calculate the scores for the five trees constructed in Section 2 for different cluster number k_0 (see Table 3). Except for the case of $k_0 = 5$, the Dumbacher's tree has the smallest misclassification error score. The trees with the next smallest score are the neighbor-joining tree and the minimum evolution tree. Since in most cases, the tree constructed by Dumbacher *et al*. is better than the other trees, we concluded that it is the optimal tree under the criterion, followed by the neighbor-joining tree and minimum evolution trees.

Note that although we use the index number for each cluster derived by the adjusted *k*-means approach to calculate the misclassification error scores, the magnitudes of these index numbers do not represent any meaning. They may be rearranged by the adjusted *k*-means approach. For

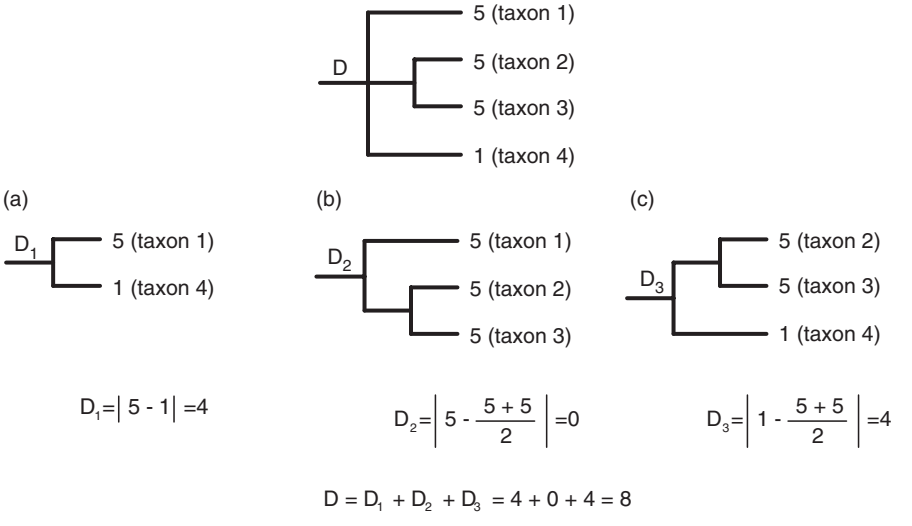

Figure 4. Decomposition of the score at node *D*.

Table 2. A rearrangement index numbers from Table 1.

k_0	2	3	4	5	6	7
A. wallacii	2	1	3	1	3	3
A. archboldi	2	1	3	1	4	4
A. albertisi	2	1	3	1	4	4
A. albertisi salvadori	2	1	3	5	2	2
A. bennettii	2	3	4	3	5	1
A. cristatus	2	3	4	3	5	1
A. crinifrons	2	1	2	4	6	5
A. tatei	1	2	1	2	1	6
A. insignis	1	2	1	2	1	7

Table 3. The misclassification error scores of the five trees for $k_0 = 2, 3, 4, 5, 6$ and 7.

Score	k_0					
(excluding the both)	2	3	4	5	6	7
Dumbacher	0	1	1.5	9	6.25	8.25
Neighbor-joining	0	2	3	6	7	10
Minimum evolution	0	2	3	6	7	10
Maximum parsimony	1	3	4.5	10.5	9.5	9
UPGMA	0	2	3	6	8.5	11.5

example, in the case of $k_0 = 5$ in Table 1, the index number is (5,5,5,1,3,3,2,4,4) can be rearranged as (1,1,1,5,4,4,2,3,3). This may lead to a different misclassification error score for the tree.

However, in spite of the fact that different index number arrangements lead to different score value, by comparing several different arrangements, we found that the results are not significantly affected by index number selection. Table 2 provides a set of rearrangement index numbers from Table 1. These two sets of index numbers lead to the same misclassification error scores shown in Table 3. Thus, we believe that the arrangement of the adjusted k-means approach will not significantly affect the score calculation.

5. Simulation result

In addition to using the owlet-nightjar real data example, we conducted a simulation study to investigate the feasibility of the adjusted k-means approach in selecting the valid tree. We use the

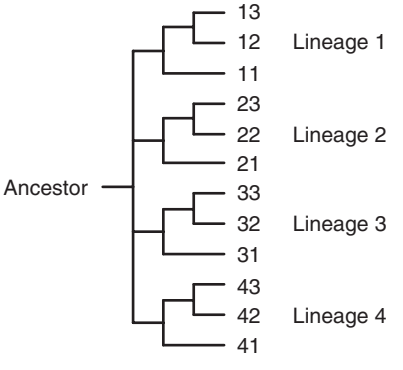

Figure 5. The generation pattern for 12 descendent sequences generated from an ancestor sequence with the Juke and Cantor model.

Juke and Cantor model [24] to generate 12 descendent sequences with a length of 100 from an ancestor sequence using the pattern in Figure 5 for different substitution rates [24].

To generate a descendant sequence from an ancestral sequence, we first set a substitution rate α, and then generate a descendant sequence with probability $1/4 + 3/4\,e^{(-4\alpha)/3}$ that the nucleotide at a site in a descendant sequence is the same as that in an ancestral sequence and with probability $1/4 + 1/4\,e^{(-4\alpha)/3}$ that the nucleotide at a site in a descendant sequence is equal to one of the three other bases from the ancestral sequence.

There are many different kinds of models. We select one of the models as an example for the simulation study. Although we do not provide the simulation study for other models, we believe the phylogenetic tree with a symmetric topology shown in Figure 5 is a good example to examine the method. However, since we do not conduct simulation studies for all models and the topology of different models is different, we may not guarantee the feasibility of the proposed method to other models. This is the limitation of the simulation study.

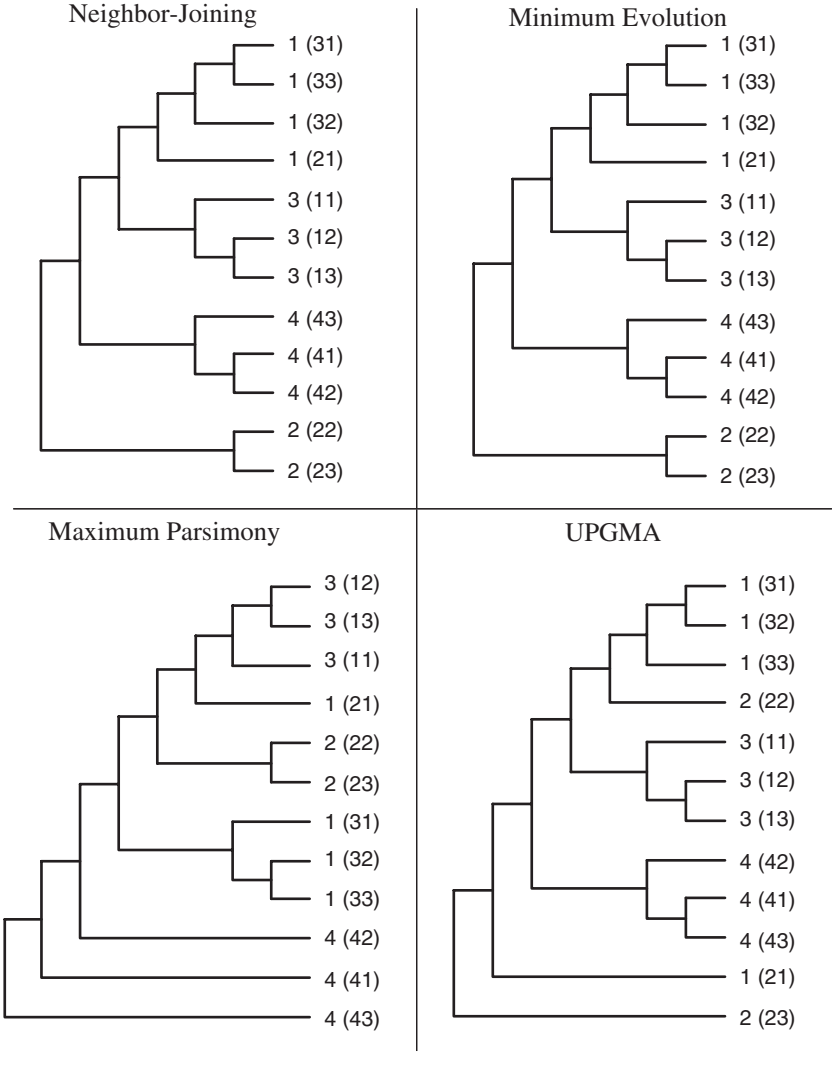

Figure 6. The four trees for the 12 descendent sequences for rate 0.01.

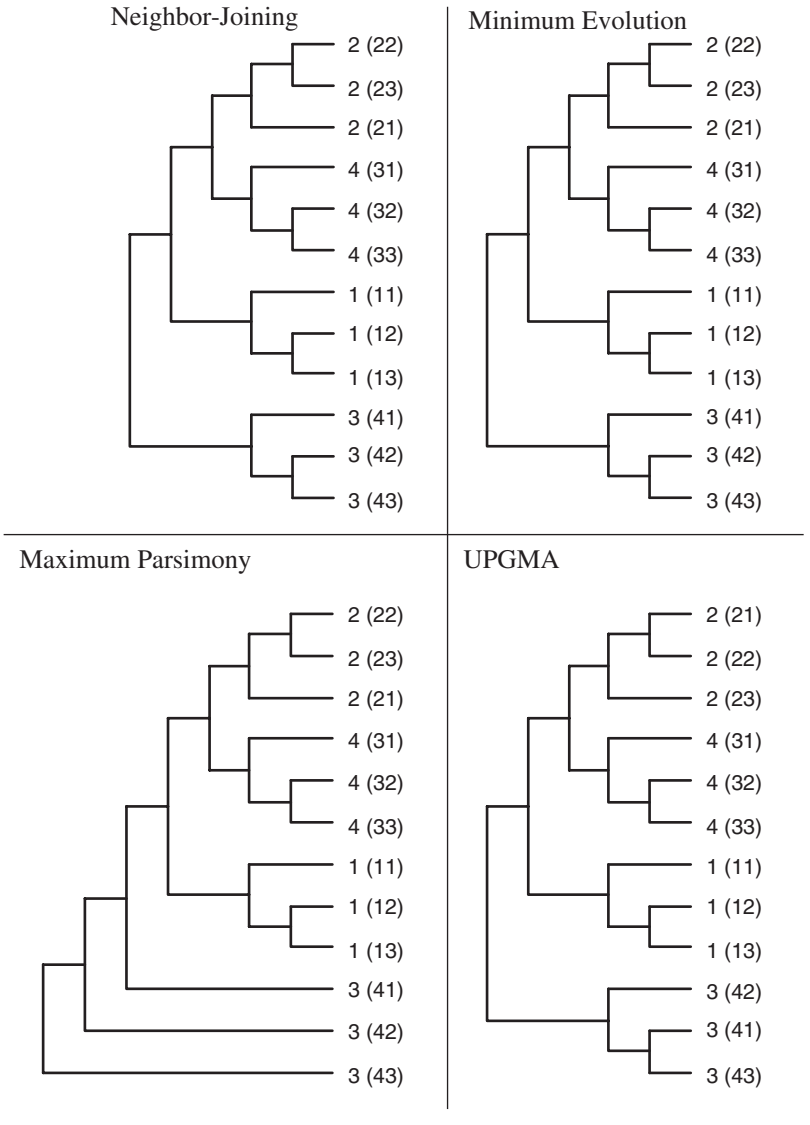

Figure 7. The four trees for the 12 descendent sequences for rate 0.1.

Table 4. The misclassification error scores of four trees for different rates of 12 sequences generated from the model in Figure 5.

	Substitution rate						
Score	0.01	0.02	0.03	0.04	0.05	0.10	0.20
Neighbor-joining	4	3.5	6	4	3.5	4	2
Minimum evolution	4	3.5	6	4	3.5	4	4
Maximum parsimony	6.67	5.67	6.5	4.33	4	6	4.5
UPGMA	6.5	7	5.17	3.5	3.5	4	4

The misclassification error scores for the four trees under $k_0 = 4$ for different substitution rates derived by the adjusted k-means approach are shown in Table 3.

We show the phylogeny trees for rate 0.01 and 0.1 in Figures 6 and 7. The number in the bracket of a tree in Figure 6 denotes the sequence number in Figure 5. The notations are the same for Figure 7. Table 4 shows that in the case of the rate being 0.01, the maximum parsimony and UPGMA trees have higher misclassification error score than the other two trees. From Figure 6, we can see that maximum parsimony and UPGMA trees have more dissimilar topologies from the topology of the tree in Figure 5 than the other two trees. For the case of rate 0.1, the maximum parsimony tree has a significantly higher misclassification error score than the other three trees. From Figure 7, among the four trees, the maximum parsimony tree has a topology most dissimilar from the tree in Figure 5. The simulation results show that the misclassification error score derived from the adjusted k-means approach can provide a good method to select a valid tree.

6. Conclusion

There are several well-known approaches for reconstructing phylogenetic trees. Since the topologies of trees constructed from different methods may be quite different, we develop a methodology to evaluate the performance of the tree from a statistical data analysis viewpoint, which is the k-means clustering approach for the categorical data associated with a misclassification error scoring algorithm. Based on this method, we can calculate the score of a tree, and then recommend trees with a smaller misclassification error score. The simulation study shows the feasibility of this method. We provide a scoring algorithm in this study, however, more than one scoring algorithm can be adopted. Discussion about the selection of a scoring algorithm is a future research topic. Since this method is to examine trees from a statistical viewpoint, it can provide an objective criterion for phylogenetic tree selection.

Acknowledgements

This study was supported by National Science Council and National Center for Theoretical Sciences in Taiwan.

References

[1] M.R. Anderberg, *Cluster for Applications*, Academic Press, London, 1973.
[2] P.C. Cosman, R.M. Gray, and R.A. Olshen, *Vector quantization: Clustering and classification trees*, J. Appl. Statist. 21 (1994), pp. 93–108.
[3] J.P. Dumbacher, T.K. Pratt, and R.C. Fleischer, *Phylogeny of the owlet-nightjars* (*Aves: Aegothelidae*) *based on mitochondrial DNA sequence*, Mol. Phylogenet. Evol. 29 (2003), pp. 540–549.
[4] D. Graur and W.H. Li, *Fundamentals of Molecular Evolution*, Sinauer Associates, Sunderland, MA, 2000.
[5] M. Hasegawa, Y. Iida, T. Yano, F. Takaiwa, and M. Iwabuchi, *Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences*, J. Mol. Evol. 22 (1985), pp. 32–38.
[6] M. Hasegawa, H. Kishino, and T. Yano, *Dating of the human–ape splitting by a molecular clock of mitochondrial DNA*, J. Mol. Evol. 22 (1985), pp. 160–174.
[7] Z. Huang, *Extensions to the k-means algorithm for clustering large data sets with categorical values*, Data Mining Knowledge Discovery 2 (1998), 283–304.
[8] M.H. Huh and Y.B. Lim, *Weighting variables in K-means clustering*, J. Appl. Statist. **36** (2009), pp. 67–78.
[9] A.K. Jain and R.C. Dubes, *Algorithms for clustering data*, Prentice Hall, New Jersey, 1988.
[10] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data – An Introduction to Cluster Analysis*, Wiley, New York, 1990.
[11] K.K. Kidd and L.A. Sgaramella-Zonta, *Phylogenetic analysis: Concepts and methods*, Am. J. Hum. Genet. 23 (1971), pp. 235–252.
[12] S. Kumar, J. Dudley, M. Nei, and K. Tamura, *MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences*, Briefings Bioinformatics 9 (2008), pp. 299–306.

[13] D.J. Lipman, S.F. Altschul, J.D. Kececioglu, *A tool for multiple sequence alignment*, Proc. Nat. Acad. Sci. USA **86** (1989), pp. 4412–4415.
[14] W.Y. Loh, *Improving the precision of classification trees*, Ann. Appl. Statist. 3 (2009), pp. 1710–1737.
[15] J. MacQueen, *Some methods for classification and analysis of multivariate observations*, in *Proceedings of the Fifth Berkeley Symposium on Mathematics Statistics and Probability*, L.M. Le Cam and J. Neyman, eds., Vol. 1, University of California Press, Statistical Laboratory of the University of California, Berkeley, 1967, pp. 281–297.
[16] M.K. Ng, M.J. Li, J.Z. Huang, and Z. He, *On the impact of dissimilarity measure in k-modes clustering algorithm*, IEEE Trans. Pattern Anal. Machine Intell. 29 (2007), pp. 503–507.
[17] N. Saitou and M. Nei, *The neighbor-joining method: A new method for reconstructing phylogenetic trees*, Mol. Biol. Evol. 4 (1987), pp. 406–425.
[18] M.R. Segal and I.B. Tager, *Trees and tracking*, Statist. Med. 12 (1993), pp. 2153–2168.
[19] W.D. Shannon and D. Banks, *Combining classification trees using MLE*, Statist. Med. 18 (1999), pp. 727–740.
[20] J. Stein, G. Kalb, K. Possinger, and K.-D. Wernecke, *Extension to the validation of classification trees*, Biometrical J. 43 (2001), pp. 107–116.
[21] K. Tamura, J. Dudley, M. Nei, and S. Kumar, *MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0*, Mol. Biol. Evol. 24 (2007), pp. 1596–1599.
[22] Y. Tateno, M. Nei, and F. Tajima, *Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species*, J. Mol. Evol. 18 (1982), 387–404.
[23] H. Wang, *Confidence intervals for the substitution number in the nucleotide substitution models*, Mol. Phylogenet. Evol. 60 (2011), pp. 472–479.
[24] H. Wang, Y.H. Tzeng, and W.H. Li, *Improved variance estimators for one- and two-parameter models of nucleotide substitution*, J. Theoret. Biol. 254 (2008), pp. 164–167.
[25] K.-D. Wernecke, K. Possinger, G. Kalb, and J. Stein, *Validating classification trees*, Biometrical J. 40 (1998), pp. 993–1005.
[26] Z. Yang, *Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods*, J. Mol. Evol. 39 (1994), pp. 306–314.

Appendix

In this section, we use five sequences which are from site 1 to site 20 of five species to illustrate the WGM and BGM calculations. The codes and calculation details are given in Tables A1–A4.

Table A1. Five sequence codes.

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A. wallacii	C	T	T	T	G	G	A	T	C	C	C	T	T	C
A. archboldi	C	T	T	T	G	G	A	T	C	C	C	T	T	C
A. albertisi	C	T	T	T	G	G	A	T	C	C	C	T	T	C
A. bennettii	–	–	T	C	G	G	A	T	C	T	C	T	C	C
A. cristatus	C	T	T	C	G	G	A	T	C	T	C	T	C	C

–, missing site.

Table A2. Calculation of WGM of a sequence X and a set $G^4 = (G_1^4, G_2^4, G_3^4, G_4^4)$.

Site		1	2	3	4	5	6	7	8	9	10	11	12
X		T	T	G	G	A	T	C	C	C	T	T	C
G^4	G_1^4	T	T	G	G	A	T	C	C	C	T	T	C
	G_2^4	T	T	G	G	A	T	C	C	C	T	T	C
	G_3^4	T	C	G	G	A	T	C	T	C	T	C	C
	G_4^4	T	C	G	G	A	T	C	T	C	T	C	C
$\sum_{i=1}^4 \phi(x_j, g_{ij})$		0	2	0	0	0	0	0	2	0	0	2	0

Table A3. Calculation for G^3.

Site			1	2	3	4	5	6	7	8	9	10	11	12
G^3	G_1^3		T	C	G	G	A	T	C	T	C	T	C	C
	$G^3 \backslash G_1^3$	G_2^3	T	C	G	G	A	T	C	T	C	T	C	C
		G_3^3	T	T	G	G	A	T	C	C	C	T	T	C
$\sum_{i=2}^{3} \phi(G_{1j}^3, G_{ij}^3)$			0	1	0	0	0	0	0	1	0	0	1	0

Table A4. BGM calculation for $\{G^2, G^3\}$.

Site		1	2	3	4	5	6	7	8	9	10	11	12
$n_1 = 2, G^2$	G_1^2	T	T	G	G	A	T	C	C	C	T	T	C
	G_2^2	T	T	G	G	A	T	C	C	C	T	T	C
$n_2 = 3, G^3$	G_1^3	T	C	G	G	A	T	C	T	C	T	C	C
	G_2^3	T	C	G	G	A	T	C	T	C	T	C	C
	G_3^3	T	T	G	G	A	T	C	C	C	T	T	C

There are two sites with missing value in the fourth sequence. By our method, we consider the sequences removing these two sites. Therefore, the sequence length is 12.

We have

$$d(X, G^4) = \frac{\sum_{j=1}^{12} \sum_{i=1}^{4} \phi(x_j, g_{ij})}{4 \times 12} = \frac{2+2+2}{4 \times 12} = 0.125.$$

Then we have

$$d\left(G_1^3, G^3 \backslash G_1^3\right) = \frac{1+1+1}{2 \times 12} = 0.125,$$

$$d\left(G_2^3, G^3 \backslash G_2^3\right) = \frac{1+1+1}{2 \times 12} = 0.125$$

and

$$d\left(G_3^3, G^3 \backslash G_3^3\right) = \frac{2+2+2}{2 \times 12} = 0.25.$$

(i) Thus, WGM for $G^3 \equiv \{G_1^3, G_2^3, G_3^3\}$

$$= d\left(G_1^3, G^3 \backslash G_1^3\right) + d\left(G_2^3, G^3 \backslash G_2^3\right) + d\left(G_3^3, G^3 \backslash G_3^3\right)$$
$$= 0.125 + 0.125 + 0.25 = 0.5.$$

(ii) BGM calculation

BGM for $\{G^2, G^3\}$

$$= [d(G_1^2, G^3) + d(G_2^2, G^3)] + [d(G_1^3, G^2) + d(G_2^3, G^2) + d(G_3^3, G^2)]$$
$$= \left[\frac{6}{3 \times 12} + \frac{6}{3 \times 12}\right] + \left[\frac{6}{2 \times 12} + \frac{6}{2 \times 12} + \frac{0}{2 \times 12}\right] = 0.833$$