Bioinformatics, 32(9), 2016, 1281–1285
doi: 10.1093/bioinformatics/btw005
Advance Access Publication Date: 6 January 2016
Discovery Note

Phylogenetics

Estimating the number and assignment of clock models in analyses of multigene datasets

Sebastián Duchêne[1,2,*], Charles S. P. Foster[1] and Simon Y. W. Ho[1]

[1]School of Life and Environmental Sciences and [2]Sydney Medical School, University of Sydney, Sydney, NSW 2006, Australia

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on 16 September 2015; revised on 3 January 2016; accepted on 4 January 2016

Abstract

Motivation: Molecular-clock methods can be used to estimate evolutionary rates and timescales from DNA sequence data. However, different genes can display different patterns of rate variation across lineages, calling for the employment of multiple clock models. Selecting the optimal clock-partitioning scheme for a multigene dataset can be computationally demanding, but clustering methods provide a feasible alternative. We investigated the performance of different clustering methods using data from chloroplast genomes and data generated by simulation.

Results: Our results show that mixture models provide a useful alternative to traditional partitioning algorithms. We found only a small number of distinct patterns of among-lineage rate variation among chloroplast genes, which were consistent across taxonomic scales. This suggests that the evolution of chloroplast genes has been governed by a small number of genomic pacemakers. Our study also demonstrates that clustering methods provide an efficient means of identifying clock-partitioning schemes for genome-scale datasets.

Availability and implementation: The code and data sets used in this study are available online at https://github.com/sebastianduchene/pacemaker_clustering_methods.

Contact: sebastian.duchene@sydney.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Evolutionary rates and timescales can be estimated from nucleotide sequences using molecular-clock models, which describe the pattern of rate variation among lineages. The various clock models share a reliance on age calibrations, but they differ in their assumptions about the number and distribution of distinct evolutionary rates (reviewed by Ho and Duchêne, 2014). For example, the strict molecular clock assumes a single rate across all lineages (Zuckerkandl and Pauling, 1962), whereas uncorrelated relaxed clocks allow branches to have distinct rates that are drawn from the same distribution (Drummond *et al.*, 2006; Rannala and Yang, 2007). There are various model-selection methods for identifying the best-fitting clock model for a dataset of interest (e.g. using marginal likelihoods; Baele *et al.*, 2013). The choice of clock model can have substantial impacts on phylogenetic estimates, particularly those of evolutionary rates and timescales.

Rates of molecular evolution often vary among lineages, but this pattern of variation can differ across sites and across genes (Gaut *et al.*, 2011; Muse and Gaut, 1997). Therefore, when complex sequence data are being analyzed, the use of multiple clock models might provide a better fit (e.g. higher marginal likelihood) than a single clock model. For example, separate clock models might be applied to different genes or codon positions.

In analyses of multigene datasets, there are usually many possible partitioning schemes. Identifying the best-fitting schemes involves two components: determining the optimal number of clusters, and assigning the genes to these clusters. In Bayesian analyses, this can be done using Bayes factors to compare different clock-partitioning schemes (Ho and Lanfear, 2010). However, such an approach is impractical when there are many candidate schemes, as is often the case for multigene or genome-scale datasets, because the statistical fit of every possible scheme would need to be assessed.

Clustering methods provide a computationally feasible means of identifying appropriate clock-partitioning schemes, by grouping subsets of the data according to their pattern of among-lineage rate variation (Duchêne *et al.*, 2014). Similar approaches are available for selecting partitioning schemes for substitution models (Frandsen *et al.*, 2015). The software ClockstaR, which was designed to identify the best-fitting clock-partitioning scheme for multigene datasets (Duchêne *et al.*, 2014), employs a *k*-medoids clustering algorithm known as partitioning around medoids (Kaufman and Rousseeuw, 2005). However, other clustering methods, such as *k*-means and Gaussian mixture modeling, have not been tested in the context of clock-model selection. One advantage of Gaussian mixture models is that they can represent the shapes of clusters flexibly by using covariance matrices. For example, they can use a diagonal covariance matrix to identify clusters with ellipsoidal shapes, such that they might have higher accuracy than *k*-medoids.

Here we test the performance of three different clustering methods for identifying the clusters of patterns of among-lineage rate variation in multigene datasets: variational inference Gaussian mixture model (VBGMM), Dirichlet process Gaussian mixture model (DPGMM), and partitioning around medoids (PAM). We evaluate these three methods using simulated data and apply them to chloroplast genome sequences from angiosperms. We find that the optimal number of clusters for these datasets range from one to three. Our results also reveal that mixture models, such as VBGMM and DPGMM, tend to detect a larger number of clusters than methods based on partitioning, such as PAM. Mixture models also appear to be more robust than PAM in that they can detect the correct number of clusters in a broader range of simulation conditions.

2 Methods

2.1 Clustering methods

We compared the performance of three different methods: VBGMM and DPGMM, as implemented in the Python module Scikit-learn v0.16 (Pedregosa *et al.*, 2012), and PAM implemented in the R package Cluster v1.15 (Maechler *et al.*, 2005). The PAM algorithm, also known as *k*-medoids, is very similar to the *k*-means algorithm. It involves randomly choosing *k* data points from the data, known as the 'medoids'. The remaining data points are assigned to their closest medoid to form *k* clusters. In the next step, the medoids are replaced by the data points that are closest to the center of each cluster, resulting in new medoids. The data points are reassigned to the new medoids. The last two steps are repeated until the medoids are the same for successive iterations. To select the optimal value of *k*, we use the Gap statistic, which uses the ratio of cluster width to distance between clusters, as a measure of goodness-of-fit (Tibshirani *et al.*, 2001). In our analyses, each cluster represents a group of genes that have similar patterns of among-lineage rate variation.

The VBGMM and DPGMM assume that the data were generated from a mixture of Gaussian probability distributions, also known as 'components', with unknown parameters. Both of these methods incorporate information about the covariance structure of the data. The most commonly used are the spherical and the diagonal covariance matrices. The spherical covariance matrix assumes that each cluster has the same variance across dimensions, resulting in spherical clusters. In contrast, in the diagonal covariance matrix the variance can differ among dimensions, such that clusters can take ellipsoidal shapes. The number of components for VBGMM is finite, so they need to be specified *a priori*. To select the optimal value, we calculate the Bayesian Information Criterion (BIC) for values of *k* from 1 to $n-1$, where *n* is the number of data points. In DPGMM, the number of components is infinite, but the number of clusters to which the data are assigned is defined by a Dirichlet process. In practice, the implementation of DPGMM requires an upper bound for the number of components, which we set as $n-1$. For both the VBGMM and the DPGMM algorithms, we used the BIC to compare the fit of diagonal and spherical covariance matrices. However, the BIC cannot be computed for PAM, such that the performance of this method cannot be assessed using this metric.

2.2 Chloroplast genome data

We obtained complete chloroplast genome sequences of angiosperms from GenBank (accession numbers in Supplementary Table S1). The advantage of analyzing genes from the non-recombining chloroplast genome is that they all share the same topology, which is an important requirement of these methods. We initially aligned all protein-coding genes using MUSCLE v3.5 (Edgar, 2004), followed by visual inspection. Three genes (*inf*A, *ycf*1 and *ycf*2) were excluded because of alignment ambiguities, leaving 76 genes for subsequent analysis, although this number varied among taxonomic groups. To reduce potential impacts of missing data, we excluded any sites in the alignment at which a gap was present for $\geq$80% of taxa.

Our initial dataset contained 183 taxa, including representatives of all major angiosperm groups. We drew subsamples to form five datasets representing different taxonomic levels: (i) angiosperms (18 taxa); (ii) eudicots (15 taxa); (iii) rosids (13 taxa); (iv) Poaceae (20 taxa); and (v) Asteraceae (7 taxa). For the Poaceae and Asteraceae datasets, some gene alignments consisted primarily of missing data, so we removed these alignments and used 61 and 74 genes, respectively, instead of the 76 genes in the complete set of gene alignments. For each of the five taxonomic datasets, we concatenated all of the genes to infer the topology using maximum likelihood in PhyML v3.1 (Guindon *et al.*, 2010) with the GTR+Γ nucleotide substitution model. We then estimated individual gene trees while constraining the tree topology to that inferred from the concatenated data, which is equivalent to optimizing the branch lengths.

Clustering algorithms typically cluster data points represented in an *n*-dimensional space. Previous studies have used individual branch lengths as dimensions in which to represent gene trees as data points (e.g. dos Reis *et al.*, 2012; Duchêne *et al.*, 2014; Duchêne and Ho, 2015). We used the same approach by treating the branch lengths as a proportion of the total tree length and using a $\log_{10}$ transformation. Our empirical data and example code are available online (https://github.com/sebastianduchene/pacemaker_clustering_methods).

2.3 Simulations

To test the performance of the clustering methods under known conditions, we first generated datasets by simulating data points using the mixture model with the highest fit according to the BIC. This involved sampling data points from the mixture of distributions inferred by the model. We also simulated data under optimal conditions for the PAM algorithm. To do this, we estimated the mean and standard deviation of each dimension for each cluster inferred using the mixture models to represent the clusters as multivariate normal distributions. We then sampled data points from these distributions. In both simulation scenarios, the simulations have the same dimensions as the chloroplast genome data described above, but they differ in the shape and spread of the clusters. We conducted 100 simulations for each of the chloroplast datasets and analyzed them using

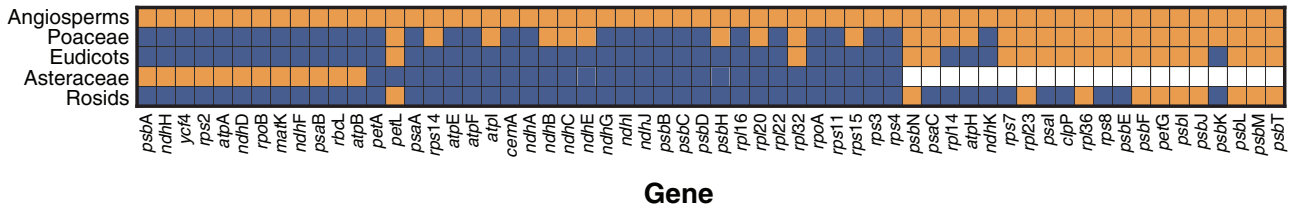

Fig. 1. Illustration of cluster assignment using the model with highest statistical fit for genes shared between the five chloroplast datasets. The rows correspond to the five datasets, and each column represents a gene. Colours indicate the cluster assignments of the genes in each dataset

all of the clustering algorithms. We analyzed the simulated datasets using the mixture models, then we selected the model with the highest statistical fit and noted the optimal value of k. We also estimated k using the Gap statistic under the PAM algorithm.

Our simulations serve two specific purposes. First, they allow us to assess the stability, or reproducibility, of the methods; that is, whether the same model and value of k is recovered for datasets generated under the same simulation conditions. Second, the simulations can be interpreted as parametric bootstrap replicates: if the inferences of the model are robust, the simulated datasets should have the same optimal model and number of clusters as those inferred for the empirical data.

3 Results

In our analyses of five chloroplast datasets, the VBGMM with a spherical covariance matrix had higher fit than the other mixture models (Table 1). Using this method, the optimal number of clusters ranged from one to three across the five datasets (Fig. 1; Supplementary Fig. S1). The PAM algorithm inferred one cluster for Poaceae, rosids, and eudicots, and two clusters for Asteraceae and angiosperms. Although the number of clusters inferred by all of the methods was similar, VBGMM with a spherical covariance matrix tended to infer the largest number of clusters. The exception was the angiosperm dataset, for which VBGMM with a spherical covariance matrix only identified one cluster, whereas the other methods identified two clusters. Importantly, for all datasets there were at least some discrepancies in the number of clusters inferred by the methods. The PAM algorithm inferred the smallest number of clusters for almost all of the datasets.

Our analyses of data simulated under mixture models showed that the algorithms with mixture models correctly identified the model used to generate the data and the number of clusters for a majority of the simulations (Table 2). In all cases, the VBGMM with a spherical covariance matrix presented the highest statistical fit for all 100 simulations. For the chloroplast data from angiosperms, Poaceae and eudicots, the estimated value of k matched the true value in all 100 simulation replicates. For rosids, the correct value of k was recovered for 97% of the simulated datasets. The mixture model fitted to the Asteraceae dataset performed the most poorly. In this case, the correct value of k was recovered for only 69% of the simulated datasets. The fact that the frequency of the optimal k is overall high for the mixture model also indicates that it is stable, yielding similar estimates for data simulated under the same conditions.

For the simulations based on mixture models, the PAM algorithm performed more poorly (Table 2). It only recovered the true value of k for the simulations using the model fitted to the angiosperms. For the other simulations, it tended to estimate a larger number of clusters, from five in the simulations using the model fitted to Poaceae to eight for the simulations under the model fitted to the rosids. The stability of this algorithm was also much lower than that of the mixture models for most datasets. For example, for the Asteraceae data, the most frequent value of k was present in 20 of the 100 simulated datasets, with many different values of k being inferred for the remaining 80 simulated datasets. This probably occurred because this dataset contains a small number of points, such that there is greater variation among simulation replicates. The simulations using the model fitted to the angiosperm data had more stable results, with a frequency of 1.00 for $k = 1$.

In our analyses of data simulated under conditions consistent with the assumptions of the PAM algorithm, we found that both mixture models and PAM recovered the correct number of clusters with a frequency of 1.00. As with the data simulated using mixture models, the VGBMM with a spherical covariance matrix had the highest statistical fit among the mixture models. Collectively, our

Table 1. Number of clusters (k) of branch-length patterns among genes in five chloroplast datasets, estimated using different clustering methods and covariance matrices

Dataset	Model	Covariance matrix	BIC	*K*
Angiosperms	VBGMM	Diagonal	10126.0	2
	VBGMM	**Spherical**	**9474.2**	**1**
	DPGMM	Diagonal	31939.1	2
	DPGMM	Spherical	20823.2	2
	PAM	–	–	2
Poaceae	VBGMM	Diagonal	9856.7	2
	VBGMM	**Spherical**	**9191.5**	**2**
	DPGMM	Diagonal	28274.2	1
	DPGMM	Spherical	18606.3	2
	PAM	–	–	1
Eudicots	VBGMM	Diagonal	8521.9	2
	VBGMM	**Spherical**	**7657.2**	**2**
	DPGMM	Diagonal	26545.8	2
	DPGMM	Spherical	17265.2	2
	PAM	–	–	1
Asteraceae	VBGMM	Diagonal	3728.3	2
	VBGMM	**Spherical**	**3465.6**	**3**
	DPGMM	Diagonal	10756.2	2
	DPGMM	Spherical	7584.7	2
	PAM	–	–	2
Rosids	VBGMM	Diagonal	7218.9	2
	VBGMM	**Spherical**	**6336.5**	**2**
	DPGMM	Diagonal	22633.8	3
	DPGMM	Spherical	14539.2	3
	PAM	–	–	1

Datasets were analyzed with the variational inference Gaussian mixture model (VBGMM), Dirichlet process Gaussian mixture model (DPGMM), and partitioning around medoids (PAM). The Bayesian information criterion (BIC) was used to compare the fit of the mixture models to each dataset, with the best-fitting model shown in bold.

Table 2. Estimated number of clusters (*k*) of branch-length patterns among genes in simulated datasets

Dataset	True *k*	k_{mixture}	Frequency of k_{mixture}	k_{PAM}	Frequency of k_{PAM}
VBGMM simulations					
Angiosperms	1	1	1.00	1	1.00
Poaceae	2	2	1.00	5	0.66
Eudicots	2	2	1.00	7	0.30
Asteraceae	3	3	0.69	7	0.20
Rosids	2	2	0.97	8	0.32
PAM simulations					
Angiosperms	1	1	1.00	1	1.00
Poaceae	2	2	1.00	2	1.00
Eudicots	2	2	1.00	2	1.00
Asteraceae	3	3	1.00	3	1.00
Rosids	2	2	1.00	2	1.00

Results are based on analyses of 100 simulations under the model fitted to each of the five chloroplast datasets. In all cases, the most frequently chosen mixture model was the VGBMM with a spherical covariance matrix (frequency of 1.00). $k_{mixture}$ is the most frequent k for analyses of the data simulated using mixture models. k_{PAM} is the most frequent k for the analyses using the PAM algorithm, with its corresponding frequency.

analyses of simulated data show that mixture models and the PAM algorithm perform well when the model used to generate the data matches that used to infer the number of clusters. However, mixture models performed well even when the data were simulated using a scenario based on PAM, such that they provide more robust estimates.

4 Discussion

We investigated the performance of three different clustering methods for grouping genes according to their pattern among-lineage rate variation. We have found that the VBGMM with a spherical covariance matrix provides the best fit among the mixture models to a range of chloroplast datasets, and our simulation study confirms the stability of this method. The PAM algorithm failed to recover the simulation conditions under VBGMM in most cases, probably because the shape of the clusters is difficult to capture using this method. In contrast, VBGMM frequently estimated the correct number of clusters irrespective of the simulation method. This differs from the results of previous studies of clustering methods for branch-length patterns, which found that the PAM algorithm appeared to perform well (Duchêne and Ho, 2014, 2015; Duchêne *et al.*, 2014). However, we reanalyzed a mammalian genome dataset from our previous study (Duchêne and Ho, 2015) and found a similar number of clusters (Supplementary Material); the most stable mixture model (DPGMM) supported seven clusters, compared with 13 using PAM in the original study. This suggests that, in empirical studies, it is important to compare the inferences from different clustering algorithms. In this study, for example, the estimated numbers of clusters for the empirical data are very similar among clustering algorithms. We find that mixture models provide a powerful alternative that can flexibly accommodate different cluster shapes. The results from these models also appear more stable under different simulation conditions, at least for the datasets analyzed here. Another advantage of these methods is that their parametric nature offers a simple framework for conducting simulations, which should be done routinely to assess the robustness of the results. Importantly, the shape of the clusters and choice of covariance structure do not necessarily have biological implications. Rather, they provide a convenient mathematical description of the cluster shapes.

The clusters identified in our analyses represent groups of genes that have similar patterns of among-lineage rate variation. All of the clustering algorithms suggest that the evolution of chloroplast genomes in angiosperms and nuclear genomes in mammals has been governed by a small number of pacemakers, each of which leads to a distinct pattern of rate variation among lineages (Ho, 2014; Snir *et al.*, 2012). This is consistent with previous findings from prokaryotes (Snir, 2014), *Drosophila*, and yeast (Snir *et al.*, 2014). Furthermore, comparing the gene clusters across our five angiosperm datasets reveals that there is some consistency in pacemakers across different taxonomic scales (Fig. 1). However, additional work will be needed to understand the biological bases of these pacemakers.

Identifying genes with similar patterns of among-lineage rate variation has important applications in phylogenetic analyses. Notably, in molecular dating studies, estimates of divergence times have been shown to be more accurate if a separate relaxed-clock model is assigned to each cluster of genes (Duchêne and Ho, 2014). Our results indicate that multigene datasets might only exhibit a small number of distinct patterns of rate variation among lineages. This has notable implications for analyses of genome-scale datasets, for which only a small number of relaxed-clock models might be sufficient to capture the key components of evolutionary rate variation. To this end, clustering methods provide a feasible and reliable alternative to more computationally demanding approaches to selecting clock-partitioning schemes for molecular dating analyses. In particular, mixture models might have better performance than the *k*-medoids and *k*-means algorithms for genomic data because they can model clusters of different shapes. Increasing the adoption of these methods will help to improve estimates of evolutionary rates and timescales from genome-scale datasets.

Funding

C.S.P.F. was supported by the Australian Award. S.Y.W.H. was supported by the Australian Research Council (grant DP110100383).

Conflict of Interest: none declared.

References

Baele,G. *et al.* (2013) Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol. Biol. Evol.*, **30**, 239–243.

dos Reis,M. *et al.* (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. R. Soc. Lond. B*, **279**, 3491–3500.

Drummond,A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLOS Biol.*, **4**, 699–710.

Duchêne,S. *et al.* (2014) ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis. *Bioinformatics*, **30**, 1017–1019.

Duchêne,S. and Ho,S.Y.W. (2014) Using multiple relaxed-clock models to estimate evolutionary timescales from DNA sequence data. *Mol. Phylogenet. Evol.*, **77**, 65–70.

Duchêne,S. and Ho,S.Y.W. (2015) Mammalian genome evolution is governed by multiple pacemakers. *Bioinformatics*, **31**, 2061–2065.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Frandsen,P.B. *et al.* (2015) Automatic selection of partitioning schemes for phylogenetic analyses using *k*-means clustering of site rates. *BMC Evol. Biol.*, **15**, 13.

Gaut,B. *et al*. (2011) The patterns and causes of variation in plant nucleotide substitution rates. *Annu. Rev. Ecol. Evol. Syst*., **42**, 245–266.

Guindon,S. *et al*. (2010) New algorithms and methods to estimate maximum likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol*., **59**, 307–321.

Ho,S.Y.W. (2014) The changing face of the molecular evolutionary clock. *Trends Ecol. Evol*., **29**, 496–503.

Ho,S.Y.W. and Duchêne,S. (2014) Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol*., **23**, 5947–5965.

Ho,S.Y.W. and Lanfear,R. (2010) Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. *Mitochondr. DNA*, **21**, 138–146.

Kaufman,L. and Rousseeuw,P.J. (2005) *Finding Groups in Data: An Introduction to Cluster Analysis*. 1st edn. Wiley, Hoboken, NJ, USA.

Maechler,M. *et al*. (2005) *Cluster Analysis Basics and Extensions*. R Statistics Package.

Muse,S.V. and Gaut,B.S. (1997) Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics*, **146**, 393–399.

Pedregosa,F. *et al*. (2012) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Rannala,B. and Yang,Z. (2007) Inferring speciation times under an episodic molecular clock. *Syst. Biol*., **56**, 453–466.

Snir,S. (2014) On the number of genomic pacemakers: a geometric approach. *Algorithms Mol. Biol*., **9**, 26.

Snir,S. *et al*. (2012) Universal pacemaker of genome evolution. *PLOS Comput. Biol*., **8**, e1002785.

Snir,S. *et al*. (2014) Universal pacemaker of genome evolution in animals and fungi and variation of evolutionary rates in diverse organisms. *Genome Biol*., **6**, 1268–1278.

Tibshirani,R. *et al*. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, **63**, 411–423.

Zuckerkandl,E. and Pauling,L. (1962) Molecular disease, evolution and genic heterogeneity. In: Kasha,M. and Pullman,B. (eds.) *Horizons in Biochemistry*. Academic Press, New York, pp. 189–225.