SPIn: Model Selection for Phylogenetic Mixtures via Linear Invariants

A. M. Kedzierska, M. Drton, R. Guigó, and M. Casanellas

1Bioinformatics and Genomics Group, Centre for Genomic Regulation (CRG) and UPF, Barcelona, Catalonia, Spain
2Department de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Barcelona, Spain
3Department of Statistics, University of Chicago
4Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

*Corresponding author: E-mail: marta.casanellas@upc.edu.

Abstract

In phylogenetic inference, an evolutionary model describes the substitution processes along each edge of a phylogenetic tree. Misspecification of the model has important implications for the analysis of phylogenetic data. Conventionally, however, the selection of a suitable evolutionary model is based on heuristics or relies on the choice of an approximate input tree. We introduce a method for model Selection in Phylogenetics based on linear Invariants (SPIn), which uses recent insights on linear invariants to characterize a model of nucleotide evolution for phylogenetic mixtures on any number of components. Linear invariants are constraints among the joint probabilities of the bases in the operational taxonomic units that hold irrespective of the tree topologies appearing in the mixtures. SPIn therefore requires no input tree and is designed to deal with nonhomogeneous phylogenetic data consisting of multiple sequence alignments showing different patterns of evolution, for example, concatenated genes, exons, and/or introns. Here, we report on the results of the proposed method evaluated on multiple sequence alignments simulated under a variety of single-tree and mixture settings for both continuous- and discrete-time models. In the simulations, SPIn successfully recovers the underlying evolutionary model and is shown to perform better than existing approaches.

Key words: linear invariants, discrete nonhomogeneous evolutionary models, phylogenetic mixtures, identifiability.

Introduction

In a probabilistic model of nucleotide evolution, the nodes of a tree $T$ are assigned random variables with four possible states in the set $\{A, C, G, T\}$. The random variables at the leaves of the tree are observed and the variables at the interior nodes are hidden. Typically, each edge, $e$, is assigned a $4 \times 4$ matrix of substitution rates between the bases. The distribution of states at the root of $T$ is denoted by $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$. The edge matrices and $\pi$ specify a continuous-time Markov chain of sequence evolution along a particular tree. Specification of an evolutionary model of suitable complexity for the nucleotide substitution process at hand is often viewed as a “preference” step in phylogenetic analysis. However, as has been emphasized in the literature (Posada and Crandall 2001; Ripplinger and Sullivan 2008), this step should be addressed with care as it can strongly impact the accuracy of the reconstructed topology and the estimates of the branch length. Inference of an appropriate evolutionary model is further challenged when the data evolved under a nonhomogeneous model (rate matrices vary across the edges) or along multiple trees (phylogenetic mixture).

Ripplinger and Sullivan (2010) show that the performance of established model selection methods depends highly on the underlying tree topology. A common practice, however, adopts a circular argument: The tree and the parameters of interest are estimated by choosing a model supported by a precomputed tree (e.g., the neighbor joining tree based on Jukes–Cantor distances). Moreover, as outlined above, available methods for selecting a model of evolution typically assume constant rate parameters at each point in time as well as a single tree topology underlying the data-generating process (e.g., Foster 2004; Huelsenbeck et al. 2004; Posada 2008). Mossel and Vigoda (2005) and Ronquist et al. (2006) discuss poor mixing of the phylogenetic Markov chain Monte Carlo (MCMC) in the presence of mixed phylogenetic signals. In this work, we propose an approach designed to deal with both nonhomogeneous and mixed data with no a priori requirement of a tree topology.

The probabilities of nucleotide bases observed at the leaves of a tree satisfy different collections of equalities depending on the evolutionary model (see, e.g., Felsenstein 2004, p. 375). Hence, as pointed out by Fu and Li (1992), Steel et al. (1992), and Felsenstein (2004), these equalities, also referred to as “linear invariants,” could potentially be used to discriminate between different models of base change. Recently, this idea was explored by Casanellas et al. (forthcoming).

In this paper, we consider discrete-time hidden Markov processes on trees assuming independence of nucleotides at different sites and the same evolutionary models for all sites. The parameters of the model are taken to be the entries of the substitution matrices that describe transitions between the nucleotides. As a result, in contrast to the continuous-time substitution models, the models defined in this paper can accommodate nonhomogeneity with different rate matrices at different lineages (see Materials...
and Methods section). According to Casanellas et al. (forthcoming), the set of probability distributions for the bases at the leaves that come from a mixture of trees under a discrete-time evolutionary model coincides with the set of distributions satisfying a certain collection of linear invariants. It is worth noting that mixtures on the same tree topology contain distributions coming from models employing discrete gamma rates ($\Gamma$) from Yang (1994) and/or invariable sites ($I$); see Steel et al. (2000) and the references therein.

Casanellas et al. (forthcoming) describe an effective algorithm to obtain the relevant linear invariants for any number of operational taxonomic units (OTUs) under some of the most widely used evolutionary models: Jukes–Cantor $JC69^*$, Kimura two-parameters $K80^*$, Kimura three-parameters $K81^*$, and the strand symmetric model SSM (Casanellas and Sullivant 2005). We use the symbol ($\subset$) to emphasize the nonhomogeneous nature of these models and to distinguish them from their respective continuous-time correspondents. The above models include as submodels the commonly used continuous-time $JC69$ (Jukes and Cantor 1969), $K80$ (Kimura 1980), $K81$ (Kimura 1981).

Also, SSM is a generalized nonhomogeneous version of HKY (Hasegawa et al. 1985), where there is an equal distribution of the pairs of bases $A, T$ and $C, G$ at each node of the tree and no assumption about a stable base distribution. All models listed above are submodels of the general Markov model GMM (Steel et al. 1994; Allman and Rhodes 2003). We have the following chain of inclusions: $JC69^* \subset K80^* \subset K81^* \subset SSM \subset GMM$. (1)

We note that the well-known general time reversible model (GTR) is a special continuous-time case of the GMM, where the rates across lineages are assumed equal.

Being strictly model specific, the collection of linear invariants provided in Casanellas et al. (forthcoming) can be used to assess whether data come from a mixture of trees evolving under one of the candidate models. We wish to stress here that phylogenetic mixtures are defined on any number of phylogenetic trees, where the tree topologies are allowed to vary. According to this definition, a model on a single tree topology, but containing different sets of parameters, is also considered a mixture.

Linear invariants of a given evolutionary model, $M$, define equalities between the joint observations of the nucleotide states at the leaves. Based on the observed data, SPIn computes the maximized log-likelihood function under $M$. Lastly, it uses the second-order Akaike Information Criterion (AIC$_2$: Akaike 1973; Sugiuara 1978) to select a model, that is, the selected model minimizes the AIC$_2$ score.

We tested SPIn on synthetic data on trees of four OTUs following the guidelines of Posada and CrANDALL (2001). The simulations were done for a wide range of parameters in the continuous-time homogeneous and discrete-time nonhomogeneous settings, for a single tree topology and along a mixture of two distributions both on the same and different tree topologies. Though at this point, the existing software packages such as jModelTest (Posada 2008), PAML (Yang 2007), PHYLIP (Felsenstein 1989), or PhyML (Guindon and Gascuel 2003) offer a larger selection of models than those included in SPIn; these methods are not consistent for phylogenetic tree mixtures. For instance, the models considered by these methods do not allow mixtures of distinct tree topologies. We demonstrate this in the Results section, where we evaluate the performance of jModelTest. Recently, Nguyen et al. (2011) used the joint patterns at the leaves to assess the fit of an inferred model and a tree to the data. In order to show that SPIn is not biased towards over-complex models, we have analyzed one of the data sets used in Nguyen et al. (2011) (see Discussion).

In addition, for a given model and a number of sequences, SPIn calculates the maximum number of trees to be considered in a mixture. As proved in section 4 of Casanellas et al. (forthcoming), mixture models with more components than a particular bound cease to be identifiable. For more on the identifiability problem, the reader is referred to, for example, Chang (1996), Stefankovic and Vigoda (2007), Allman et al. (2010).

Materials and Methods

Let $\tau$ be a set of tree topologies on a set of $n$ OTUs. We consider nucleotide substitution models assuming that all nucleotides in a DNA sequence evolve independently and under the same evolutionary process. A (discrete-time) hidden Markov process of nucleotide substitution on a rooted tree topology $T$ on $n$ leaves is given by specifying a root distribution $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ and substitution matrices $S^e$ for each edge $e$ in $T$. The entries of $S^e$ are the conditional probabilities $S^e_{xy}$ that a nucleotide $x$ at the parent node of edge $e$ is substituted by nucleotide $y$ at the descendant node of edge $e$. The form of the substitution matrices and the root distribution specify the evolutionary model $M$ under consideration. For example, $M = JC69^*$ if $\pi$ is uniform and the substitution matrices have only one free parameter in the off-diagonal entries (with row sums equal to one). In particular, the models considered here do not assume homogeneity of rate matrices. Indeed, if the transition matrices were of type $S^* = \exp(t_eQ^e)$ (i.e., as a continuous-time model where $Q_e$ is a rate matrix), then lack of relations between $S^e_1$ and $S^e_2$ for two edges $e_1$ and $e_2$ of the same tree allows for $Q^e_1 \neq Q^e_2$. In other words, continuous-time nonhomogeneous models are a special case of our models (imposing transition matrices to have exponential form and equating rate matrices). We refer the reader to Allman and Rhodes (2007) for the precise description of the models considered in this paper (i.e., $JC69^*$, $K80^*$, $K81^*$, and SSM). As shown in Allman and Rhodes (2007) and Casanellas et al. (forthcoming), the placement of the root does not play a role, thus, we assume that the trees are unrooted.

Given a model $M$, a tree topology $T$, a root distribution $\pi$, and a set of substitution matrices $S = \{S^e\}_{e \in E(T)}$, let $p_M^T(\pi, S)$ be the probability vector determining the probability distribution at the leaves of $T$ under the Markov process. The entries of $p_M^T(\pi, S)$ are thus the $4^n$ probabilities $p_{M^T_{\text{cluster}}}(\pi, S)$ of observing each nucleotide pattern
(x₁, ..., xₙ) at the leaves of T under parameters (π, S).

As shown in Casanellas et al. (forthcoming), the vectors p_T^M(π, S) satisfy certain linear equations irrespective of the tree topology T and the parameters (π, S). We call them the “linear invariants of the model” M (see also Felsenstein 2004, chapter 22). For example, it is easy to see that

\[ p_{T,\alpha_1...\alpha_p}(π, S) = p_{T,\alpha_1...\alpha_p}(π, S), \]
\[ p_{T,\alpha_1...\alpha_p}(π, S) = p_{T,\alpha_1...\alpha_p}(π, S), \]
\[ p_{T,\alpha_1...\alpha_p}(π, S) = p_{T,\alpha_1...\alpha_p}(π, S), \]

are three linear invariants of JC69*. The exhaustive list of linear invariants of the above models can be easily computed. Moreover, the set of distributions satisfying these equations can be proven to coincide with the set of distributions coming from some mixture of trees under the same model. A distribution from a mixture on m trees in the set \( T \) under a model \( M \) is a joint distribution \( p = (p_{\alpha_1...\alpha_p}, p_{\alpha_1...\alpha_p} \ldots, p_{\alpha_1...\alpha_p}) \) such that

\[ p = \sum_{i=1}^{m} \alpha_i p_{T_i}^M(π, S), \]

where \( T_i \in \tau, \sum_{i=1}^{m} \alpha_i = 1 \)

(cf. e.g., Stefankovic and Vigoda 2007; Matsen et al. 2008). Note that adopting this definition assumes that the model \( M \) is the same for all \( T \). The mixing coefficients, \( \alpha_i \), represent the percentage of sites that evolved along \( T_i \). Note that under this definition, the model (Markov process) \( M \) is the same for all \( T \). Further background on phylogenetic mixtures can be found in Gascuel and Guindon (2007).

Casanellas et al. (forthcoming) give an efficient algorithm to compute the invariants of the models treated here. The linear invariants are binomials—each of them is of the form \( p_T^M = p_T^M \). Due to the nesting of models as seen in equation (1), all invariants can be obtained recursively.

Selecting a model based on biological data requires a statistical assessment of the vanishing of the linear invariants for each model. Let \( H^M \) be the linear space formed by all distributions satisfying the linear invariants for the model M. For the models considered here, \( H^M \) is defined by equalities among pairs of entries of \( p_T^M(π, S) \). Hence, the maximum likelihood estimate is unique, that is, given data D, there exists a unique point \( θ^* \in H^M \), for which the likelihood function \( L(θ, M) = \text{Prob}(D | θ, M) \) attains its maximum for \( θ \in H^M \). To score the models, we use a variant of the AICc, which includes a small sample correction along with the penalty for model complexity:

\[ \text{AIC}_c = -2 \log(L(θ^*, M)) + 2d + \frac{2(d+1)}{L - d - 1}, \]

where L is the sample size (alignment length) and d is the dimension of the linear space \( H^M \). The dimension of the \( H^M \) equivalent models can be explicitly calculated (Casanellas et al. forthcoming): \( \dim(H^{JC69}) = \frac{1}{2}n^2 - 3 + 2n - 2 + \frac{1}{2} \), \( \dim(H^{K80}) = 2n^2 - 3 + 2n - 2, \dim(H^{K81}) = 4n - 1, \dim(H^{K82}) = 2n - 1 \). The number of invariants for each model is \( 4^n \) minus its dimension.

The model selected by SPIn is the one that minimizes \( \text{AIC}_c \). For ranking purposes, the output of the algorithm includes the ratios of normalized Akaike weights

\[ w_i = e^{-\frac{1}{2}Δ_i}, \quad Δ_i = \text{AIC}_{c,i} - \min_j(\text{AIC}_{c,j}), \]

and \( \text{AIC}_{c,i} \) is the \( \text{AIC}_c \) score of a model \( M_i \). SPIn is a C++ package available at http://genome.crg.es/cgi-bin/phylomod_sel/AlgModelSelection.pl (last accessed 22 Nov 2011). The results reported in this paper use the \( \text{AIC}_c \), though Bayesian Information Criterion is available as an option in SPIn (Schwarz 1978; Burnham 2004). Moreover, ongoing work includes an implementation of an MCMC algorithm to deal with large and sparse data sets.

## Results

### Data

In order to assess the performance of SPIn in recovering the underlying model from \{JC69*, K80*, K81*, SSM*\}, we simulated multiple sequence alignments on an unrooted quartet tree following the design of Posada and Crandall (2001). Specifically, we used the quartet tree space proposed by Huelsenbeck (1995), which is defined by a pair of branch-length parameters (a, b), where a determines the length of the internal branch and two peripheral branches taken from different clades and b gives the length of the two remaining branches. Parameters a and b, representing the expected number of substitutions per site, were varied from 0.01 to 0.75 in increments of 0.02 (compare also figs. 2 and 3 in Huelsenbeck 1995).

We simulated 100 gap-free multiple sequence alignments of 300, 1,000, and 3,000 sites for every point \((a, b)\) on the grid. The alignments were generated either under a single tree topology or mixtures of two trees (see below). We then computed the fraction of alignments for which the true model with a minimal set of parameters was selected from the pool of candidate models. In graphical displays, a point \((a, b)\) is colored black if there was a 100% successful recovery. White points on the grid correspond to a 0% recovery, and the values in between the two extrema are represented in a gray scale.

We used the “evolver” program from the package PAML (Yang 2007) to generate the data under the continuous-time homogeneous JC69 and K80 models. We assumed a transition transversion ratio of 2 for K80 (κ = 4). In order to generate the data under the discrete hidden Markov process, we used the relation \( br = -\frac{1}{2} \log \det(S^*) \) between the branch length br and the determinant of the substitution matrix \( S^* \). We created a Matlab package, which we refer to as genNon−h, for simulating alignments under the discrete-time models (available at http://genome.crg.es/cgi-bin/phylomod_sel/AlgModelSelection.pl, last accessed 22 Nov 2011).

We performed a number of tests on the data simulated under different parameter and model choices. In this paper, we present a selection of the results that allowed
F1. Plots of the fraction of correctly identified models for multiple sequence alignments of length 300 or 1,000 generated on a single quartet tree (ST) under JC69, K80, K81, JC69*, K80*, and K81*; SPIn: (a) and (c); jModelTest: (b) and (d). The parameters vary in the quartet tree space: (a, b) of Huelsenbeck (1995). Fractions are displayed in gray scale ranging from 0% in white to 100% in black. Corresponding average recovery rates are given in table 1a and b.

the comparison of the performance of SPIn to that of jModelTest. The remaining data are provided in the supplementary material, Supplementary Material online.

Single Tree
We generated data on a single four-taxon tree topology and the tree space as defined above. The resulting set of data-generating distributions is denoted by ST. The results of running SPIn under the JC69 and K80 models are shown in figure 1a. It can be seen that already for alignments as short as 300 nt, the recovery is close to perfect across the entire tree space.

The average recovery for 300 nt alignments was 99.9% and 97.7% and improved to 99.7% and 99.8% for length 1,000; see table 1a. Figure 1c shows the recovery of the discrete-time JC69*, K80*, and K81* models also to be high even for short alignments. The average recovery taken over the tree space and alignments of length 1,000 was 99.7%, 96.5%, and 96.8% for JC69*, K80*, and K81*, respectively (see table 1b).

Two-Tree Mixtures
For the purpose of testing model recovery using SPIn on phylogenetic mixtures, we considered two-tree mixtures on both the same and different quartet tree topologies.

First, we generated continuous-time mixture data on the same tree topology by allowing two gamma classes in the evolver package from PAML. The pattern of model recovery under the JC69 and K80 along these two-tree mixtures is almost identical to that for a single tree; see table 1a.
Next, we tested the performance on two-tree mixture data under the discrete-time hidden Markov models JC69*, K80*, and K81*. Multiple sequence alignments were simulated by choosing a pair of tree topologies on four sequences, \( \tau_1 \) and \( \tau_2 \), with branch lengths fixed for \( \tau_1 \) and the branch lengths of \( \tau_2 \) varying over the tree space described above. We denote by “mixture on the same topology (MST),” the data-generating distributions obtained by assuming the same tree topology \( \tau_1 = \tau_2 \) and by “mixture on distinct topologies (MDT),” the distributions given by two different topologies \( \tau_1 \neq \tau_2 \). We considered two sets of branch lengths for \( \tau_1 \) in the MST and MDT data sets:

1. 0.11 for the inner branch length and two opposite peripheral branches, 0.61 for the remaining branches with a fraction of \( \lambda = 0.3 \) sites evolving on \( \tau_1 \) (0.7 evolved on \( \tau_2 \)). This selection comprises the MST\(_1\) and MDT\(_1\) data sets.

2. 0.31 for the inner branch length and two opposite peripheral branches, 0.41 for the remaining branches with a fraction of \( \lambda = 0.5 \) randomly selected sites coming from the alignment evolved on \( \tau_1 \). The corresponding data sets are denoted by MST\(_2\) and MDT\(_2\).

In concordance with the single-tree case, the recovery of the JC69* model for the MST data exceeds 99% for alignments as short as 300 nt, irrespective of the choice of the parameters. See figure 2a and table 1c for the results on 300 and 1,000 nt. As expected, it remained true for the MDT data (fig. 2c, table 1c), where the model was correctly identified at the 99% level in all data sets: 300 nt simulated for MDT\(_2\) and 3,000 nt for both MDT\(_1\) and MDT\(_2\). At length 300 nt, the K80* model was recovered on average in 54% of the cases for the MST\(_1\) (supplementary fig. IIb, Supplementary Material online) and 48% of the cases for the MST\(_2\) data set. Similarly lowered is the performance for the K81* at
FIG. 2. Plots of the fraction of correctly identified models for multiple sequence alignments of lengths 300 and 1,000 along two-tree mixtures on quartet trees on the MST under JC69* and K80*; SPIn: (a); jModelTest: (b); and on MDT under JC69* for 300 and 3,000 nt; SPIn: (c); jModelTest: (d). The parameters vary in the quartet tree space: (a, b) of Huelsenbeck (1995). Fractions are displayed in gray scale ranging from 0% in white to 100% in black. Corresponding average recovery rates are given in table 1c and d.

The reason for this relatively low performance is the high number of parameters allowed in the (*) models due to the nonhomogeneity assumption. Thus, longer sequence alignments are required when using the AICc criterion.

For all models and their parameter choices, the recovery exceeded 99% when the alignment length was 3,000 nt (supplementary figs. II and III, Supplementary Material online).

Larger Trees on Real-Life Topologies
In order to investigate the performance of SPIn when the number of OTUs is larger, we ran the tests on multiple sequence alignments simulated on two topologies inferred for real-life sets of species. As before, evolver package (PAML) was used to generate 100 multiple sequence alignments in the following settings: continuous-time JC69 model with three discrete Γ-rate classes and length 5,000 on the 9-taxon Drosophila tree (Pollard et al. 2006; Clark et al. 2007) and HKY (Hasegawa et al. 1985) model with four Γ-rate classes, transition/transversion ratio of \( \kappa = 2 \), nucleotide frequencies of \( \pi_A = \pi_C = 0.1, \pi_G = \pi_T = 0.4 \), and length 1,000 along the 12-taxon T12b yeast tree (Marcet-Houben and Gabaldón 2009); see figure 3. In both cases, the parameter α of the Γ distribution was set to 0.5.

Though the tree of Drosophila has fewer sequences than the fungal tree, its branches are shorter, which in practice
will lead to fewer different observed nucleotide patterns at the leaves. Therefore, in this case, we simulated longer alignments of 5,000 nt. In both data sets, in 100% of the cases, SPIn recovered the model the data were sampled from.

In addition, we tested the performance on the ten-taxon primate tree model obtained from Fujita et al. (2010) under continuous-time J69 and K80 three- and four-tree mixture models (supplementary material, Supplementary Material online). Since primate species are closely related, the resulting tree will have short length and presents challenges for model inference. We found that for 100% model recovery, the required alignments lengths were on average 30,000. Although this number might appear large, it is not unrealistic with the growing availability of complete genomes.

The method presented here is based on the nucleotide patterns recorded at the leaves of the tree; therefore, it is better suited for more diverged trees. In practice, including distinct clades or an outgroup (as seen in the trees used here for simulations) will significantly improve the accuracy of model recovery.

**Comparison to Existing Methods**

Existing phylogenetic packages, as mentioned in the Introduction, rely on a similar model testing principle: An initially inferred phylogeny is used to select a model for subsequent tree inference. We decided to compare the performance of SPIn with that of jModelTest, which is a popular package designed specifically for model selection.

We are aware that jModelTest was not created to deal with the discrete-time mixture data. In order to allow maximum comparability between the two methods, we chose the following settings for the command line version of jModelTest: AIC criterion with the option of five models, enabled invariant sites, and two gamma classes (−AICc, −8 S −1 −g 2). This ensured a fair comparison as the pool of models activated in jModelTest was contained within the models we considered. Although jModelTest supports neither discrete-time Markov models nor mixtures on a single or different tree topologies, we found it interesting to evaluate its performance on this type of data.

The results for the continuous-time J69 and K80 models on a single tree are shown in figure 1b and table 1d. The average model recovery was 60% and did not depend on the length of the alignments. In comparison to the continuous-time models, the average recovery for the ST data under the discrete-time models dropped to 56% for the J69* model, 37% for the K80*, and 49% for the K81* models. Interestingly, the recovery rate was found to be worse with an increase of the alignment length from 300 to 1,000, see figure 1d and table 1d.

The same trend, though with a slightly lower impact, was found for two-mixture data on the same topology, MST1, under the K80* model. The mean recovery decreased from 41% in the 300-nt data set to 37% for 1,000 nt (fig. 2b). The average detection for both MST2 and MDT2 data sets under J69* dropped with an increase of the alignment length from 67% and 65% (300 nt) to 56% and 45% (1,000 nt), respectively (table 1d and fig. 2b and d). The average model recovery on the MDT1 data set was found to be the lowest (45%) among all the tests for J69* model.

Since SPIn was designed specifically to deal with phylogenetic mixtures and nonhomogeneous data, the method outperforms jModelTest for the alignments generated under discrete-time models on single and mixture of trees. This result is due to the fact that, as proved in Casanellas et al. (forthcoming), the linear invariants are strictly model specific and derived from the properties of the nucleotide substitution matrices as opposed to the exponential rate matrices.

In species tree reconstruction, an assumption of a single tree topology is reasonable, and the data are usually composed of the alignments of single copy homologous genes. However, though the tree topology remains the same, the branches might differ in lengths along the alignment, thus, it becomes a mixture model. Unless the inference is performed on each block separately allowing for nonhomogeneity of the rates at different lineages, this fact is not accounted for by the existing methods. In such instances, as shown in the above comparison, an incorrect model is very likely to be selected and this in turn may confound the tree inference. Though it was found that in some instances an approximated model might allow for recovering the species topology, the parameter estimates will not be correct. It can be seen in the results presented here that the methods that account for mixtures increase the reliability of the results.
Discussion

We introduced a novel approach to selection of an evolutionary model in phylogenetics. SPIn uses linear invariants defining the spaces of all phylogenetic mixtures under a given model. The structure of a phylogenetic mixture model, for instance, the number of components and tree topologies, is allowed to vary freely. Although more statistical work is required to better address scenarios where a large number of sequences must be handled simultaneously, tests on simulated data coming from a single tree as well as mixtures of trees suggest that SPIn correctly identifies the underlying model in cases that proved difficult for existing methods.

Another issue regarding some of the existing methods is the tendency to select complex models. For instance, as found by Nguyen et al. (2011), in the analysis of 6,171 protein-coding regions, the GTR class of models was selected in more than 70% of the cases (see table 3 of Nguyen et al. 2011). This was also the case for the protein-coding DNA alignment (PF02724) from the PANDIT database (Whelan et al. 2006) analyzed by these authors. As shown in the quoted paper, the tree topology inferred under the GTR + I + Γ model is incongruent with the accepted phylogeny. However, using JC69 + I + Γ, the tree topology is correctly recovered. We have analyzed this data set and the model selected by SPIn is in fact JC69*. This provides evidence that SPIn does not always choose most complex models for real data sets.

We propose using SPIn as a first inference step to discriminate between mixtures on JC69*, K80*, K81*, SSM. If, for instance, the data supported JC69*, further analysis could address the question of whether an unmixed JC69, JC69+Γ, or JC69+Γ+I fits the data better. One could also investigate the number of different tree topologies that should be taken into account.

In the current version of the program, gaps and ubiquitous characters are removed from the alignment. Note that the number of invariants for each model is 4^n minus its dimension. Although this number is exponential in n, the implementation of SPIn uses only the invariants containing the patterns observed in the data. As the length of the alignment is not exponential in n, the algorithm in fact uses a subset of invariants. This approach significantly speeds up the algorithm. Current implementation limits the maximum number of input species in SPIn to 21. However, an ongoing work is to extend this number to increase applicability to the modern real-life analyses.

Here, we demonstrated good performance for up to ten species with up to 100,000 sites when using AICc. Ongoing work on sampling-based statistical inference aims at extending the applicability of SPIn to larger number of species. This said, the patterns and rates of evolution that characterize functional elements depend on their location within the genome, the G+ C content of the region, synonymous codon site selection (features addressed by accounting for mixture models), and tend to be clade specific (Pollard et al. 2010). In large studies, we recommend grouping the sequences and performing the selection on such subsets. Also, in order to deal with incomplete or new genomes, future release of SPIn will include methods to deal with highly sparse data and short alignments.

An attractive feature of SPIn is its speed. Irrespective of the model considered, the time to run SPIn on a dual-core Intel machine (2.40 GHz) with 48 GB of RAM on a multiple sequence alignment of four OTUs of length 300 was on average 0.014 s, 0.020 s for length 3,000, and 0.177 s for ten-taxon multiple sequence alignments of length 30,000 nt. As a comparison, in the latter case jModelTest took 6 min 28 s.

In addition, one of the future goals is to provide the user with valuable information on whether the data evolved along a mixture on different tree topologies, a mixture on the same tree topology, or from a single tree. We expect that phylogenetic invariants (although in this case, they cease to be linear) can be used for this purpose. At this point, however, only a few invariants are known for these cases (see, e.g., Allman et al. 2010), and further development of mathematical tools is required. Finally, we are working on expanding the set of available models. This work includes the Algebraic Time reversible and the Stable Base Distribution models (Allman and Rhodes 2006) and covarion model (Tuffley and Steel 1998; Galtier 2001).

Supplementary Material

Supplementary figures I, II, and III are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

M.C. and A.M.K. were partially supported by Spanish government MTM2009-14163-C02-02 and Generalitat de Catalunya 2009 SGR 1284. A.M.K. would like to credit the FPI grant BES-2007-16623 of Ministerio de Ciencia e Innovación. R.G. was supported by grants RD07/0067/0012, BIO2006-03380, and CSD2007-00050 from the Spanish Ministry of Science.

References


