

SPIn: model selection for phylogenetic mixtures via linear invariants

A. M. Kedzierska^{1,2}, M. Drton⁴, R. Guigó^{1,3}, M. Casanellas^{2,**}

1. Bioinformatics and Genomics Group, Center for Genomic Regulation

Barcelona Biomedical Research Park (PRBB)

c/ Dr. Aiguader, 88, 08003 Barcelona

Tel. +34 93 316 01 10; Fax +34 93 316 00 99

2. Dpt. Matemàtica Aplicada I, Universitat Politècnica de Catalunya

Avda. Diagonal 647 08028-Barcelona. Spain.

3. Universitat Pompeu Fabra

4. Department of Statistics, University of Chicago, IL

5734 S. University Ave Chicago, IL 60637, U.S.A

***corresponding author: 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 non-homogeneous phylogenetic data consisting of multiple sequence alignments showing different patterns of evolution, e.g. 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.

Keywords: linear invariants, discrete non-homogeneous evolutionary models, phylogenetic mixtures, identifiability.

Introduction

In a probabilistic model of nucleotide evolution the nodes of a tree \mathcal{T} are random variables with 4 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 4×4 matrix of substitution rates between the bases. The distribution of states at the root of \mathcal{T} is denoted by $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$. The edge matrices and π 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 ‘pre-inference’ 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 non-homogeneous 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 interests are estimated by choosing a model supported by a pre-computed 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 non-homogeneous 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. (2011).

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 that transitions between the nucleotides. As a result, in contrast to the continuous-time substitution models, the models defined in this paper can accommodate non-homogeneity with different rate matrices at different lineages (see Materials and methods section). According to Casanellas et al.

(2011), 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 (Γ) from Yang (1994) and/or invariable sites (I); see Steel et al. (2000) and the references therein.

Casanellas et al. (2011) 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 2-parameters K80*, Kimura 3-parameters K81*, and the strand symmetric model SSM (Casanellas and Sullivant (2005)). We use the symbol (*) to emphasize the non-homogeneous 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 non-homogeneous 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 (Allman and Rhodes, 2003; Steel et al., 1994). We have the following chain of inclusions:

$$\text{JC69}^* \subset \text{K80}^* \subset \text{K81}^* \subset \text{SSM} \subset \text{GMM}.$$

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. (2011) can be used to assess whether data comes 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, \mathcal{M} , define equalities between the joint ob-

servations of the nucleotide states at the leaves. Based on the observed data, SPIn computes the maximized log-likelihood function under \mathcal{M} . Lastly, it uses the second order Akaike Information Criterion (AIC_c : Akaike, 1973; Sugiura, 1978) to select a model, that is, the selected model minimizes the AIC_c score.

We tested SPIn on synthetic data on trees of 4 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 non-homogeneous 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, 1993) 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. (2011), mixture models with more components than a particular bound cease to be identifiable. For more on the identifiability problem the reader is referred to e.g. Chang (1994), Stefankovic and Vigoda (2007), Allman et al. (2010).

Material and Methods

Let τ 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 \mathcal{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 \mathcal{T} . The entries of S^e are the conditional probabilities $S^e_{x,y}$ that a nucleotide x at the parent node of edge e is substituted by nucleotide y at the descending node of edge e . The form of the substitution matrices and the root distribution specify the evolutionary model \mathcal{M} under consideration. For example, $\mathcal{M} = \text{JC69}^*$ if π is uniform and the substitution matrices have only one free parameter in the off-diagonal entries (with row sums to equal to one). In particular, the models considered here do not assume homogeneity of rate matrices. Indeed, if the transition matrices were of type $S^e = \exp(t_e Q^e)$ (that is, 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, e_2 of the same tree allows for $Q^{e_1} \neq Q^{e_2}$. In other words, continuous-time non-homogeneous 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 (that is, JC69^* , K80^* , K81^* and SSM). As shown in Allman and Rhodes (2007) and Casanellas et al. (2011) the placement of the root does not play a role, thus we assume that the trees are unrooted.

Given a model \mathcal{M} , a tree topology \mathcal{T} , a root distribution π and a set of substitution matrices $S = \{S^e\}_{e \in E(\mathcal{T})}$, let $p_{\mathcal{T}}^{\mathcal{M}}(\pi, S)$ be the probability vector determining the probability distribution at the leaves of \mathcal{T} under the Markov process. The entries of $p_{\mathcal{T}}^{\mathcal{M}}(\pi, S)$ are thus the 4^n probabilities $p_{T, x_1 \dots x_n}^{\mathcal{M}}(\pi, S)$ of observing each nucleotide pattern (x_1, \dots, x_n) at the leaves of \mathcal{T} under parameters (π, S) . As shown in Casanellas et al. (2011), the vectors $p_{\mathcal{T}}^{\mathcal{M}}(\pi, S)$ satisfy certain linear equations irrespective of the tree topology \mathcal{T} and the parameters (π, S) . We call them the *linear invariants of the model \mathcal{M}* (see also Felsenstein, 2004, chapter 22). For example, it is easy to see that

$$p_{\mathcal{T}, AA \dots A}^{\mathcal{M}}(\pi, S) = p_{\mathcal{T}, CC \dots C}^{\mathcal{M}}(\pi, S), p_{\mathcal{T}, CC \dots C}^{\mathcal{M}}(\pi, S) = p_{\mathcal{T}, GG \dots G}^{\mathcal{M}}(\pi, S), p_{\mathcal{T}, GG \dots G}^{\mathcal{M}}(\pi, S) = p_{\mathcal{T}, TT \dots T}^{\mathcal{M}}(\pi, S)$$

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 τ under a model \mathcal{M} is a joint distribution $p = (p_{AA\dots A}, p_{AA\dots C}, \dots, p_{TT\dots T})$ such that

$$p = \sum_{i=1}^m \alpha_i p_{\mathcal{T}_i}^{\mathcal{M}}(\pi_i, S_i), \text{ where } \mathcal{T}_i \in \tau, \sum_{i=1}^m \alpha_i = 1$$

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

Casanellas et al. (2011) 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 $p_{\mathcal{T},X}^{\mathcal{M}} = p_{\mathcal{T},Y}^{\mathcal{M}}$. Due to the nesting of models as seen in (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 $H^{\mathcal{M}}$ be the linear space formed by all distributions satisfying the linear invariants for the model \mathcal{M} . For the models considered here, $H^{\mathcal{M}}$ is defined by equalities among pairs of entries of $p_{\mathcal{T}}^{\mathcal{M}}(\pi, S)$. Hence, the maximum likelihood estimate is unique, that is, given data D there exists a unique point $\hat{\theta} \in H^{\mathcal{M}}$ for which the likelihood function $\mathcal{L}(\theta, \mathcal{M}) = \text{Prob}(D \mid \theta, \mathcal{M})$ attains its maximum for $\theta \in H^{\mathcal{M}}$. To score the models, we use a variant of the *AIC* which includes a small sample correction along with the penalty for model complexity:

$$AIC_c = -2\log(\mathcal{L}(\hat{\theta}, \mathcal{M})) + 2d + \frac{2d(d+1)}{L-d-1},$$

where L is the sample size (alignment length) and d is the dimension of the linear space $H^{\mathcal{M}}$. The dimension of the $H^{\mathcal{M}}$ equivariant models can be explicitly calculated (Casanellas et al., 2011): $\dim(H^{\text{JC69}^*}) = \frac{1}{3}2^{2n-3} + 2^{n-2} + \frac{1}{3}$, $\dim(H^{\text{K80}^*}) = 2^{2n-3} + 2^{n-2}$, $\dim(H^{\text{K81}^*}) = 4^{n-1}$, $\dim(H^{\text{SSM}}) = 2^{2n-1}$. The number of invariants for each model is 4^n minus its dimension.

The model selected by SPIn is the one that minimizes AIC_c . For ranking purposes, the output of the algorithm includes the ratios of normalized Akaike weights

$$\frac{w_i}{w_j}, \text{ where } w_i = \frac{e^{-\frac{1}{2}\Delta_i}}{\sum_i e^{-\frac{1}{2}\Delta_i}}, \quad \Delta_i = AIC_{c,i} - \min(AIC_{c,i})$$

and $AIC_{c,i}$ is the AIC_c score of a model \mathcal{M}_i . SPIn is a C++ package available at

http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgModelSelection.pl.

The results reported in this paper use the 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 Figures 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 sets 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 grey 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 ($\kappa = 4$). In order to generate the data under the discrete hidden Markov process, we used the relation $br = -\frac{1}{4} \log \det(S^e)$ between the branch length br and the determinant of the substitution matrix S^e . 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/phylo_mod_sel).

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 the comparison of the performance of SPIn to that of *jModelTest*. The remaining data is provided in the Supplementary Material.

Single tree

We generated data on a single 4-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 1(a). It can be seen that already for alignments as short as 300nt, the recovery is close to perfect across the entire the tree space.

The average recovery for 300nt alignments was 99.9% and 97.7% and improved to 99.7% and 99.8% for length 1,000; see Table 1(a). Figure 1(c) 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 1(b)).

Two tree mixtures

For the purpose of testing model recovery using SPIn on phylogenetic mixtures, we considered 2-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 2 gamma classes in the *evolver* package from *PAML*. The pattern of model recovery under the JC69 and K80 along these 2-tree mixtures is almost identical to that for a single tree; see Table 1(a).

Next, we tested the performance on 2-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 4 sequences, τ_1 and τ_2 , with branch lengths fixed for τ_1 and the branch lengths of τ_2 varying over the tree space described above. We denote by *MST* (*mixture on the same topology*) the data-generating distributions obtained by assuming the same tree topology $\tau_1 = \tau_2$ and by *MDT* (*mixture on distinct topologies*) the distributions given by two different topologies $\tau_1 \neq \tau_2$. We considered two sets of branch lengths for τ_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 τ_1 (0.7 evolved on τ_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 τ_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 300nt, irrespective of the choice of the parameters. See Figure 2(a) and Table 1(c) for the results on 300nt and 1,000nt. As expected, it remained true for the *MDT* data (Figure 2(c), Table 1(c)), where the model was correctly identified at the 99% level in all data sets: 300nt simulated for *MDT*₂ and 3,000nt for both *MDT*₁ and *MDT*₂. At length 300nt the K80* model was recovered on average in 54% of the cases for the *MST*₁ (Fig.II(b), Supplementary material) and 48% of the cases for the *MST*₂ data set. Similarly lowered is the

performance for the K81* at the alignment length of 300: 47% for the MST_1 and 37% for the MST_2 (Fig. II(c), Supplementary material).

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

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

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 5,000nt. In both data sets SPIn recovered the model the data was sampled from in 100% of the cases.

In addition, we tested the performance on the 10-taxon primate tree model obtained from Fujita et al. (2010) under continuous-time JC69 and K80 3- and 4- tree mixture models (Supplementary material). 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 to 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_c criterion with the option of 5 models, enabled invariant sites and two gamma classes (`-AICc -s 5 -i -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 JC69 and K80 models on a single tree are shown in Figure 1(b) and Table 1 (d). 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 JC69* 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 1(d) and Table 1(d).

The same trend, though with a slightly lower impact, was found for 2-mixture data on the same topology, *MST*₁, under the K80* model: the mean recovery decreased from 41% in the 300nt data set to 37% for 1,000nt (Fig. 2(b)). The average detection for both *MST*₂ and *MDT*₂ data sets under

JC69* dropped with an increase of the alignment length from 67% and 65% (300nt) to 56% and 45% (1,000nt), respectively (Tab. 1(d) and Fig. 2(b), 2(d)). The average model recovery on the MDT_1 data set was found to be the lowest (45%) among all the test for JC69* model.

Since SPIn was designed specifically to deal with phylogenetic mixtures and non-homogeneous 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. (2011), 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 is 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 non-homogeneity 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, 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. While 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 $GTR + I + \Gamma$ model is incongruent with the accepted phylogeny. However, using $JC69 + I + \Gamma$, 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 $JC69$, $JC69 + \Gamma$, or $JC69 + \Gamma + I$ 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 10 species with up to 100,000 sites when using AIC_c . 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 which 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 2-core Intel machine (2.40GHz) with 48 GB of RAM on a multiple sequence alignment of 4 OTUs of length 300 was on average 0.014s, 0.020s for length 3000 and 0.177s for 10-taxon multiple sequence alignments of length 30000nt. As a comparison, in the latter case *jModelTest* took 6m28s.

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 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).

1 Supplementary material

The alignments used in the paper and the `genNon-h` package for simulating multiple sequence alignments under the discrete-time models are available at http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgModelSelection.pl.

2 Acknowledgements

Marta Casanellas and Anna Kedzierska were partially supported by Spanish government MTM2009-14163-C02-02 and Generalitat de Catalunya 2009 SGR 1284. Anna Kedzierska would like to credit the FPI grant BES-2007-16623 of Ministerio de Ciencia e Innovación. Roderic Guigó

was supported by the Ministerio de Educación y Ciencia with grant number CSD2007-00050.

References

- Akaike, H. Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *Second International Symposium on Information Theory. Budapest*, pages 267–281. Akadémiai Kiadó, 1973.
- Allman, E. and Rhodes, J. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 186(2):113–144, 2003.
- Allman, E. and Rhodes, J. Phylogenetic invariants for stationary base composition. *Journal of Symbolic Computation*, 41(2):138 – 150, 2006. Computational Algebraic Statistics.
- Allman, E. and Rhodes, J. Phylogenetic invariants. In Gascuel, O. and Steel, M., editors, *Reconstructing Evolution*. Oxford University Press, 2007.
- Allman, E., Petrovic, S., Rhodes, J., and Sullivant, S. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 2010.
- Burnham, K. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, November 2004.
- Casanellas, M. and Sullivant, S. The strand symmetric model. In Pachter, L. and Sturmfels, B., editors, *Algebraic Statistics for computational biology*, chapter 16. Cambridge University Press, 2005.
- Casanellas, M., Fernandez-Sanchez, J., and Kedzierska, A. The space of phylogenetic mixtures of equivariant models. in preparation, 2011. URL <http://www.ma1.upc.edu/~casanellas/mixtures.pdf>.

- Chang, J. Inconsistency of Evolutionary Tree Topology Reconstruction Methods When Substitution Rates Vary Across Characters, 1994.
- Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., and Pollard, D. A. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167):203–18, November 2007.
- Felsenstein, J. PHYLIP– Phylogeny Inference Package (version 3.2). *Cladistics*, pages 164–166, 1993.
- Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates, Inc., 2004.
- Foster, P. Modeling compositional heterogeneity. *Systematic Biology*, 53(3):485–495, 2004.
- Fu, Y. and Li, W. Construction of linear invariants in phylogenetic inference. *Mathematical Biosciences*, 109(2):201–228, 1992.
- Fujita, P., Rhead, B., Zweig, A., Hinrichs, A., Karolchik, D., Cline, M., Goldman, M., Barber, G., Clawson, H., Coelho, A., et al. The ucsc genome browser database: update 2011. *Nucleic Acids Research*, 39:1–7, 2010.
- Galtier, N. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, 18(5):866–873, 2001.
- Gascuel, O. and Guindon, S. Modelling the variability of evolutionary processes. In Gascuel, O. and Steel, M., editors, *Reconstructing Evolution*. Oxford University Press, 2007.
- Guindon, S. and Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52:696–704, 2003.
- Hasegawa, M., Kishino, H., and Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

- Huelsenbeck, J. Performance of phylogenetic methods in simulation. *Systematic Biology*, 44(1): 17–48, 1995.
- Huelsenbeck, J., Larget, B., and Alfaro, M. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution*, 21(6):1123–1133, 2004.
- Jukes, T. and Cantor, C. Evolution of protein molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism*, volume 3, pages 21–132. Academic Press, 1969.
- Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.
- Kimura, M. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the USA*, 78(1):454–458, 1981.
- Marcet-Houben, M. and Gabaldón, T. The tree versus the forest: The fungal tree of life and the topological diversity within the yeast phylome. *PLoS One*, 4:8, 2009.
- Matsen, F., Mossel, E., and Steel, M.
- Mossel, E. and Vigoda, E. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–2209, 2005.
- Nguyen, M., Klaere, S., and von Haeseler, A. MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. *Molecular Biology and Evolution*, 28(1):143–52, Jan 2011.
- Pollard, D. A., Iyer, V. N., Moses, A. M., and Eisen, M. B. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS genetics*, 2(10): e173, October 2006.
- Pollard, K., Hubisz, M., Rosenbloom, K., and Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

- Posada, D. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, 25(7): 1253–1256, 2008.
- Posada, D. and Crandall, K. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Molecular Biology and Evolution*, 18(6):897–906, 2001.
- Ripplinger, J. and Sullivan, J. Does choice in model selection affect maximum likelihood analysis? *Systematic Biology*, 57(1):76–85, 2008.
- Ripplinger, J. and Sullivan, J. Assessment of Substitution Model Adequacy Using Frequentist and Bayesian Methods. *Molecular Biology and Evolution*, 27(12):2790–2803, 2010.
- Ronquist, F., Larget, B., Huelsenbeck, J. P., Kadane, J. B., Simon, D., and van der Mark, P. Comment on “phylogenetic mcmc algorithms are misleading on mixtures of trees”. *Science (New York, N.Y.)*, 312(5772):367; author reply 367, April 2006.
- Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, March 1978.
- Steel, M., Hendy, M., Székely, L., and Erdős, P. Spectral analysis and a closest tree method for genetic sequences. *Appl. Math. Lett.*, 5(6):63–67, 1992.
- Steel, M., Székely, L., and Hendy, M. Reconstructing trees when sequence sites evolve at variable rates. *Journal of Computational Biology*, 1(2):153–163, 1994.
- Steel, M., Huson, D., and Lockhart, P. Invariable sites models and their use in phylogeny reconstruction. *Systematic Biology*, 49:225–232, 2000.
- Stefankovic, D. and Vigoda, E. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Systematic Biology* 56 (1), pages 113–124, 2007.
- Sugiura, N. Further analysts of the data by akaike’ s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1):13–26, 1978.

Tuffley, C. and Steel, M. Modeling the covarion hypothesis of nucleotide substitution. *Mathematical biosciences*, 147(1):63–91, January 1998.

Whelan, S., de Bakker, P., Quevillon, E., Rodriguez, N., and Goldman, N. Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees, 2006.

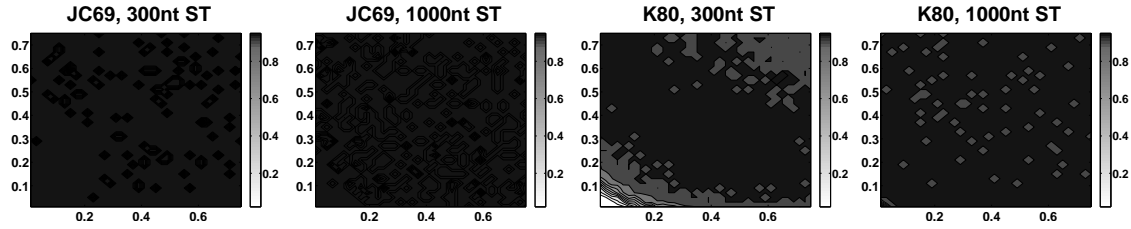
Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314, 1994.

Yang, Z. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

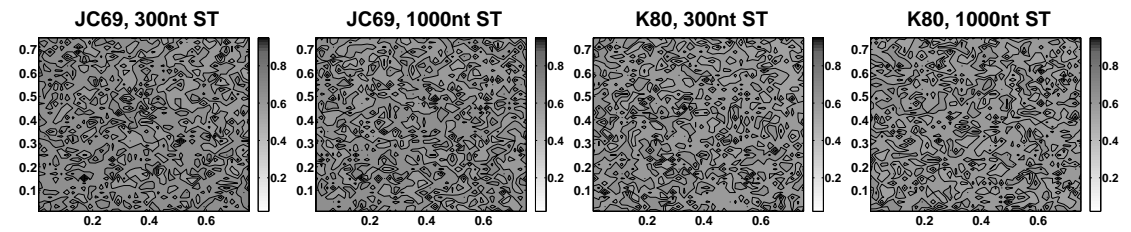
Table 1: Average recovery rate of the continuous-time (JC69, K80) and discrete-time models (JC69*, K80* and K81*) across the quartet tree space (a, b) . *ST*: single tree under continuous and discrete-time models; Γ : single tree under continuous-time model with 2 gamma rates; *MST*₁, *MST*₂: 2-tree mixture on the same topology under discrete-time models; *MDT*₁, *MDT*₂: 2-tree mixture on different topologies under discrete-time models (see Results).

(a) SPIn					(b) SPIn						
Model	JC69	JC69	K80	K80	Model	JC69*	JC69*	K80*	K80*	K81*	K81*
Length	300	1,000	300	1,000	Length	300	1,000	300	1,000	300	1,000
ST	0.999	0.997	0.977	0.998	ST	0.999	0.997	0.684	0.965	0.561	0.968
Γ	0.999	0.998	0.940	0.998							

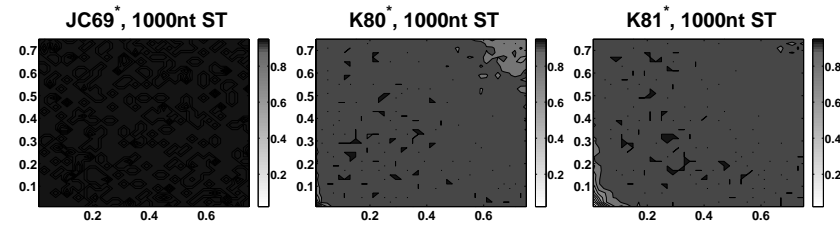
(c) SPIn					(d) <i>jModelTest</i>							
	Length	JC69*	K80*	K81*	Model	JC69	K80	JC69*	K80*	K81*		
<i>MST</i> ₁	300	0.999	0.538	0.470	Length	300	1000	300	1,000	1,000	1,000	1,000
<i>MST</i> ₂	300	0.998	0.478	0.370	ST	0.666	0.653	0.629	0.624	0.564	0.375	0.493
<i>MST</i> ₁	1000	0.997	0.935	0.590								
<i>MST</i> ₂	1000	0.997	0.929	0.965	Model	JC69*	K80*	JC69*	K80*	K81*		
<i>MST</i> ₁	3000	0.997	0.994	0.999	Length	<i>MST</i> ₂	<i>MST</i> ₂	<i>MDT</i> ₁	<i>MDT</i> ₂			
<i>MST</i> ₂	3000	0.994	0.993	0.998	Length	300	1,000	300	1,000	3,000	300	3,000
<i>MDT</i> ₁	300	0.999	0.575	0.492								
<i>MDT</i> ₂	300	0.999	0.502	0.379								
<i>MDT</i> ₁	1000	0.997	0.957	0.984								
<i>MDT</i> ₂	1000	0.998	0.952	0.977								
<i>MDT</i> ₁	3000	0.996	0.997	0.999								
<i>MDT</i> ₂	3000	0.998	0.997	0.999								



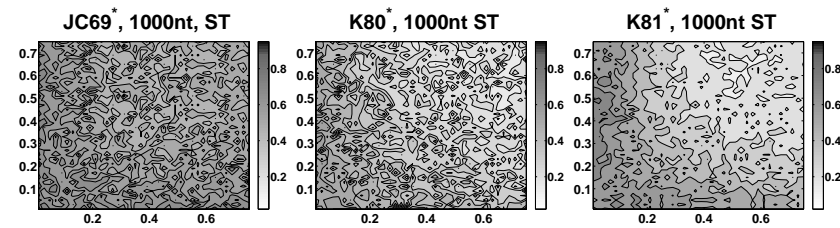
(a) SPIn



(b) *jModelTest*

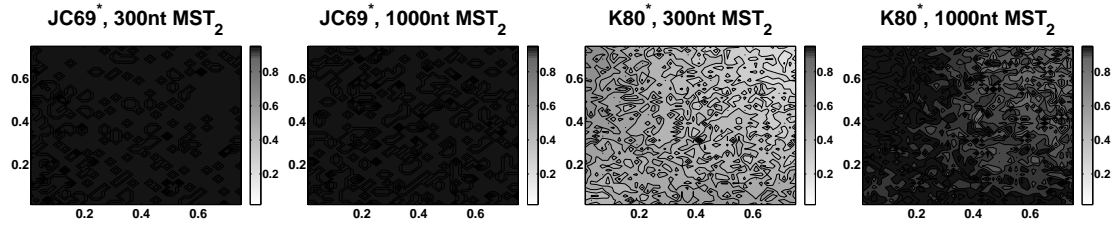


(c) SPIn

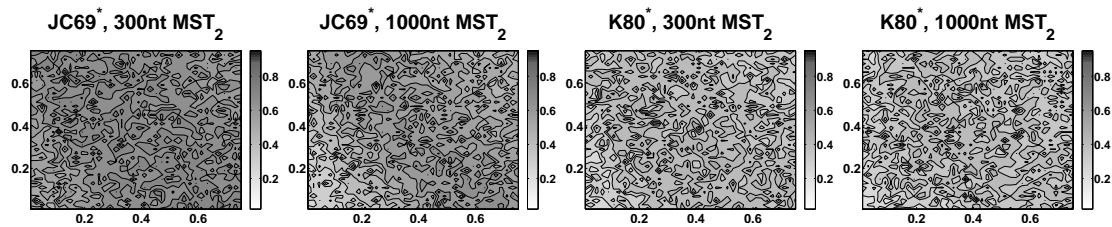


(d) *jModelTest*

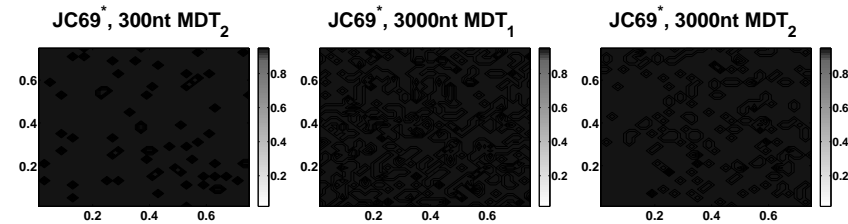
Figure 1: Plots of the fraction of correctly identified models for multiple sequence alignments of length 300 or 1,000 generated on a single quartet tree (*ST*) under JC69, K80, K81, JC69*, K80* and K81*; SPIn: (a), (c); *jModelTest*: (b), (d). The parameters vary in the quartet tree space: (*a, b*) of Huelsenbeck (1995). Fractions are displayed in grey-scale ranging from 0% in white to 100% in black. Corresponding average recovery rates are given in Table 1(a) and (b).



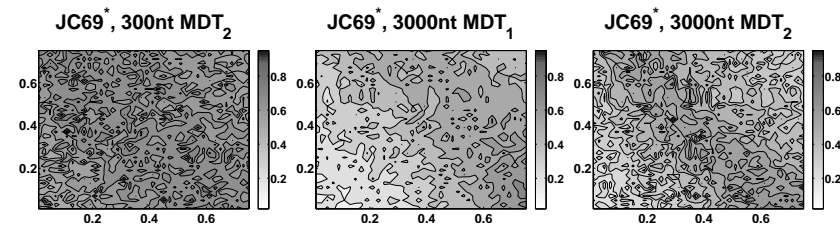
(a) SPIn



(b) *jModelTest*



(c) SPIn



(d) *jModelTest*

Figure 2: Plots of the fraction of correctly identified models for multiple sequence alignments of lengths 300 and 1,000 along 2-tree mixtures on quartet trees on the same tree topology (*MST*) under JC69* and K80*; SPIn: (a); *jModelTest*: (b); and on different tree topologies (*MDT*) under JC69* for 300nt and 3000nt; SPIn: (c); *jModelTest*: (d). The parameters vary in the quartet tree space: (a,b) of Huelsenbeck (1995). Fractions are displayed in grey-scale ranging from 0% in white to 100% in black. Corresponding average recovery rates are given in Table 1(c) and (d).

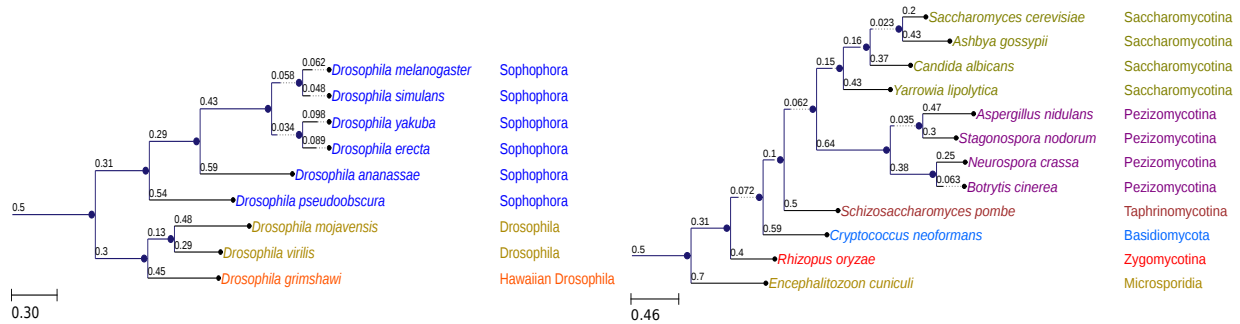


Figure 3: Phylogenetic trees used for simulations: 9-taxon drosophila (Pollard et al., 2006; Clark et al., 2007) and 12-taxon fungal tree T12b (Marcet-Houben and Gabaldón, 2009).