Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees

M. Casanellas and J. Fernández-Sánchez

Department of Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Barcelona, Spain

An attempt to use phylogenetic invariants for tree reconstruction was made at the end of the 80s and the beginning of the 90s by several researchers (the initial idea due to Lake [1987] and Cavender and Felsenstein [1987]). However, the efficiency of methods based on invariants is still in doubt (Huelsenbeck 1995; Jin and Nei 1990). Probably because these methods only used few generators of the set of phylogenetic invariants. The method studied in this paper was first introduced in Casanellas et al. (2005) and it is the first method based on invariants that uses the “whole” set of generators for DNA data. The simulation studies performed in this paper prove that it is a very competitive and highly efficient phylogenetic reconstruction method, especially for nonhomogeneous models on phylogenetic trees.

Introduction

Since the introduction of phylogenetic invariants by Cavender and Felsenstein (1987), Lake (1987), and Evans and Speed (1993), several attempts to give a generating set of polynomial phylogenetic invariants have been made (see, e.g., Steel et al. 1993; Ferretti and Sankoff 1995), but it has not been until recently that algebraic geometers have managed to find them all (Allman and Rhodes 2004a; Casanellas and Sullivant 2005; Sturmfels and Sullivant 2005). Methods based on invariants have already proved to be useful in comparative genomics (Sankoff and Blanchette 1999). However, a perception seems to have developed that invariants are inefficient, in the technical sense of requiring long sequences to correctly infer a phylogenetic tree, compare (Jin and Nei 1990) and (Huelsenbeck 1995). Although Huelsenbeck (1995) showed the inefficiency for the use of Lake’s invariants alone, no method using all invariants had been proposed at that point, and the question for invariants-based methods in general was never investigated. Note that Lake’s method of invariants only used 2 phylogenetic invariants of degree one among the 795 generators of the set of polynomial invariants of a quartet tree for the Kimura 2-parameter model (Garcia and Porter 2005). But as Felsenstein explained, invariants are worth more attention for “what they might lead to in the future” (Felsenstein 2003). This future may be soon, because the studies of this paper show a method based on invariants, which is indeed promising. Recently, other methods based on a large set of invariants have also been considered (Eriksson 2005; Kim et al. 2006).

Phylogenetic invariants are relationships satisfied by the expected pattern frequencies occurring in sequences evolving along a given tree topology T under an evolutionary model. More precisely, if t is the set of d model parameters on T and \( p_s(t) \) is the probability of observing the pattern \( s \) at the leaves of \( T \), by letting \( t \) vary on an open subset of \( \mathbb{R}^d \), the probability vector \( p(t) = (p_{\lambda A A A}(t), \ldots, p_{TT \ldots T}(t)) \) defines a subset \( S_T \) of dimension \( \leq d \) of \( \mathbb{R}^d \). A “phylogenetic invariant” is a real-valued continuous function \( f(x) \) on \( \mathbb{R}^d \) such that \( f(p) = 0 \) for any \( p \in S_T \), but not for all the points on the subset \( S_T \) determined by another tree topology \( T' \). Essentially, the equations \( f(p) = 0 \) are satisfied for pattern frequencies arising from any model parameters on a fixed tree, so they might be used for recovering the tree topology.

In practice, the vector of observed pattern frequencies \( \hat{p} \) obtained from an alignment of sequences for \( n \) taxa with enough data should approximate \( p(t) \) for some set of parameters \( t \) on a tree topology \( T \). In other words, \( \hat{p} \) should be a point close to the subset \( S_T \), so if \( f \) is an invariant for the topology \( T \), one should have \( f(\hat{p}) \) very close to 0. As the tree topology is identifiable via invariant-based methods (Allman and Rhodes 2006), using phylogenetic invariants for tree reconstruction is a consistent method (see Hagedom and Landweber 2000; Felsenstein 2003). A practical introduction to the theory of invariants can be found in the book by J. Felsenstein (2003), Chapter 22, whereas the book by Pachter and Sturmfels (2005) provides a beautiful insight into the applications of algebraic statistics (and in particular polynomial phylogenetic invariants) to computational biology.

There are 2 major motivations for using phylogenetic invariants in tree reconstruction. One of them is the prohibitive computational expense of a full maximum likelihood estimation of a tree, its edge lengths, a base distribution, and a rate matrix. The other is that the evolutionary models underlying the theory of invariants allow for nonhomogeneous mutation. Indeed, it is known that, for some biological data sets, different rate matrices should be allowed in different lineages. Thus, it is essential to have at our disposal phylogenetic methods for reconstructing trees admitting nonhomogeneous models (Galtier and Gouy 1998; Yang and Yoder 1999).

In this paper, a phylogenetic reconstruction method that uses polynomial phylogenetic invariants (introduced in Casanellas et al. 2005) is studied and tested for quartet trees evolving under the Kimura (1981) 3-parameter model of nucleotide substitution. Actually, we consider an “algebraic” Kimura model: the parameters of the model are the entries of the substitution matrices on the edges (and not a single rate matrix together with edge lengths). Hence, the model is nonhomogeneous—because it allows different rate matrices among the edges—but it is stationary (and the distribution of the bases is uniform), and we always assume that all sites are independent and identically distributed (i.i.d. hypotheses). We performed simulation studies to test its efficiency. One of our approaches to evaluate the performance and efficiency of the method is taken from

Key words: invariants, phylogenetic inference, molecular evolution, Markov process, phylogenetic tree, nonhomogenous.

E-mail: martacasanellas@upc.edu.
Huelsenbeck (1995) so that a large portion of the tree space is examined to get a general idea of how the algorithm performs. We present the results obtained for sequences of length 100 up to 10,000 base pairs.

We also checked the performance of the method on simulated data from nonhomogeneous models by carrying out a comparison to Neighbor-Joining algorithm (Saitou and Nei 1987), maximum likelihood algorithm (Felsenstein 1981), and an algorithm for a nonhomogeneous model from PAML (Yang 1997) for sequences generated under a Kimura (1980) 2-parameter model and different rate matrices along different tree branches.

Results

Homogeneous Data

The performance of the invariants method studied here on homogeneous Kimura models can be seen in the figure 1. Using the approach of J.P. Huelsenbeck (1995) for quartet trees, we considered 2 branch-length parameters \( a \) and \( b \) and simulated data for each pair of lengths. Parameter \( a \) assigns the branch length to the internal branch and 2 opposite peripheral branches, and parameter \( b \) assigns the branch length to the 2 remaining branches. Parameters \( a \) and \( b \) were varied from 0.01 to 0.75 in increments of 0.02, and 1,000 alignments were simulated for each couple \((a, b)\) (see Huelsenbeck 1995, Figures 2 and 3 or fig. I, Supplementary Material online, for a clear picture of this space of parameters). The simulated trees evolve under the Kimura (1981) 3-parameter model with a fixed rate matrix of the following form along the tree \((- = - \alpha - \beta - \gamma)\):

\[
\begin{pmatrix}
A & C & G & T \\
\gamma & \alpha & \beta & \\
\alpha & \beta & \gamma & \\
\beta & \gamma & & 
\end{pmatrix}
\]

Figure 1 shows the efficiency of the method considered in this paper for a rate matrix with parameters \( \gamma = 0.1, \alpha = 3.0, \) and \( \beta = 0.5 \) (hence, a 5:1 transition:transversion bias) and for nucleotide sequences of lengths 100, 500, 1,000, and 10,000. See also figures III and IV, Supplementary Material online, for studies performed with other rate matrices. This figure is to be compared with those shown in Figure A2 of (Huelsenbeck 1995), (corresponding to phylogenetic inference for sequences generated under a Kimura (1980) 2-parameter model of substitution with 5:1 transition:transversion bias). Though this is of course a biased comparison because our method admits nonhomogeneous data and Kimura 3-parameter model, it is worth noticing that even on data from homogeneous simulations our method outperforms many of the methods considered there. In particular, it is clearly better than Lake’s invariant method (which is not surprising because Lake’s method only used linear invariants). At least for sequences of 500 base pairs or more, our method performs better than Neighbor-Joining—referred as minimum evolution (Kimura\(_1\)) in Huelsenbeck (1995), Figure A2. Notice also that the shape of the 95% isocline for lengths around 500 base pairs more is quite different from the corresponding shapes in the methods tested by Huelsenbeck (see his Figure 7): for large values of \( a \) (near 0.7), the performance of our method of invariants does not drop drastically (as it does for most methods considered there). Therefore, for these values, our invariants method outperforms all the methods studied in Huelsenbeck (1995). As it can be seen in figures III and IV, Supplementary Material online, it seems that the method performs better for small transition: transversion ratios.

For length 1,000, the efficiency of the invariants method considered here is similar to that obtained for lengths \( \geq 10,000 \) in many of the methods tested in Huelsenbeck (1995). From this it can be inferred that, in order to reconstruct the correct tree, much less data is needed in the invariants method presented here than in many other methods (contrary to what was thought until now in Hagedorn and Landweber 2000).

Nonhomogeneous Data

We tested the invariants method studied in this paper on data simulated according to a nonhomogeneous Kimura model by comparing it with other methods.

Comparison with Neighbor-Joining

First of all, we compared the performance of the invariants method presented here with Neighbor-Joining (the algorithm of Saitou and Nei 1987) using Kimura (1981) 3-parameter distance. As it can be seen in figure 2, considering certain nonhomogeneous sets of simulated data, we found that the invariants method is more efficient than Neighbor-Joining. Indeed, we simulated data on an unrooted quartet tree evolving under the Kimura 3-parameter model where different rate matrices at each edge were chosen (see fig. 3), and we studied the efficiency of the method when varying the length of nucleotide sequences. In this case, the mean of correctly reconstructed trees for the invariants method is 90.2%, whereas the mean for Neighbor-Joining is 84%. It is worth pointing out that the use of Kimura distance is still justified in the nonhomogeneous setting, so that it is fair to compare both methods. For this particular tree, the maximum likelihood algorithm for the Kimura 3-parameter model reconstructed the tree correctly almost all times, so we do not include the results here.

Comparison with Maximum Likelihood

We compared the invariants method with 2 versions of maximum likelihood: the usual maximum likelihood for Kimura (1980) 2-parameter and nonhomogeneous maximum likelihood method developed in the package PAML (Yang 1997) for Kimura 2-parameter model (namely, the option nhomo). This last method allows different transition/transversion ratio in different tree branches. To perform this comparison, we simulated data according to the tree in figure 3 evolving under the Kimura 2-parameter model. When \( \varepsilon = 1 \), the tree is homogeneous and we studied the efficiency of the 3 methods as \( \varepsilon \) increases up to 9. Figure 4 summarizes the results relative to the comparison between
our method, the nonhomogeneous method in PAML for Kimura 2-parameter model (option nhomo=2 in the baseml control file) and the maximum likelihood homogeneous in PAML for Kimura 2-parameter (option nhomo=0). It shows the percentage of correctly reconstructed trees for each value of parameter $e$. As expected, the homogeneous maximum likelihood algorithm becomes less efficient as $e$ increases. The use of a method assuming homogeneity on data produced in accord with a nonhomogeneous model is, of course, not recommended as such model misspecification is known to lead to unreliable results. Already for $e = 5$, the invariants method overtakes the homogeneous maximum likelihood algorithm. As it is deduced from figure 4, our method performs worse than the nonhomogeneous algorithm in PAML. However, it is worth noticing that in this test we generated data according to Kimura 2-parameter model and the invariants method presented here was developed under Kimura 3-parameter model. Similar results are obtained when the data evolve under Kimura 3-parameter model (see figure II, Supplementary Material online).

Methods

The phylogenetic reconstruction method used in this paper is based on phylogenetic invariants and was first introduced by the first author and L.D. Garcia and S. Sullivant in Casanellas et al. (2005).

The taxa are given by an alignment of $n$ DNA sequences of length $N$. A Markov process along an unrooted binary
tree of \( n \) species is assumed, and we consider that all sites are independently and identically distributed (i.i.d. hypotheses). The parameters of the model we are considering are the entries of the substitution matrices of a Kimura (1981) 3-parameter model (we should rather speak of an “algebraic” Kimura 3-parameter model, according to the book by Pachter and Sturmfels 2005, Chapters 1 and 4). Note that in the usual Kimura 3-parameter model, a rate matrix \( Q \) is fixed and common to the whole tree and the substitution matrix is the exponential \( e^{Qt} \) (for some parameter \( t \) representing time). However, in our method, we do not make use of rate matrices \( Q \)—we only use the substitution matrices—so, as we will see later, the rate matrices might vary among the edges. It is worth pointing out that, as we consider a Kimura 3-parameter model along all branches, the uniform distribution of base composition holds for the whole tree (stationarity hypothesis).

**Phylogenetic Invariants**

Sturmfels and Sullivant (2005) gave an explicit description of the generators of the set of polynomial phylogenetic invariants \( I(T) \) for an arbitrary tree evolving under a “group-based model\(^{12} \).” For the Kimura 3-parameter model on an unrooted 4-taxon tree, the ideal of phylogenetic invariants has 8,002 minimal generators (see the Web page of Garcia and Porter 2005, Casanellas et al. 2005 and the discussion in Sturmfels and Sullivant 2005, Section 7 to see why a smaller subset of invariants does not suffice). According to the results in Sturmfels and Sullivant (2005), we produced this generating set for an unrooted tree with 4 leaves under the Kimura 3-parameter model. This requires doing a Fourier transform (or Hadamard conjugation) on the vector of probabilities \( \{p_A...p_{sT}\} \), and the phylogenetic invariants are then described as binomials in the Fourier coordinates.

**Algorithm**

Our tree reconstruction algorithm performs the following tasks. Given 4 aligned DNA sequences \( s_1, s_2, s_3, \) and \( s_4 \), it first computes the observed relative frequencies of each pattern for the topology \( (s_1, s_2), (s_3, s_4) \) on an unrooted quartet tree. Then it transforms these relative frequencies to Fourier coordinates. From this, we compute the Fourier coordinates in the other 2 possible topologies for unrooted trees with 4 species. We then evaluate all phylogenetic invariants for the Kimura 3-parameter model in the Fourier coordinates of each tree topology. We call \( s_f^T \) the absolute value of this evaluation for the polynomial \( f \) and tree topology \( T \). From the values \( \{ s_f^T \} \), we produce a score for each tree topology \( T \), namely \( s(T)=\sum_f s_f^T \). The algorithm then
chooses the topology that corresponds to the minimum score. The code was written in PERL and is available upon request. For an alignment of 4 sequences of 1,000 nt, it takes 0.35 s on a single 3.0-GHz processor.

Software Used

The simulations of this study were obtained using the program Seq-Gen v1.3.2 (Rambaut and Grassly 1997). We used the algorithms implemented in the package APE (Paradis et al. 2006) v1.8-3 of R (R Development Core Team 2005) v2.1.1 to compute Kimura (1981) 3-parameter distance and perform Neighbor-Joining algorithm (Saitou and Nei 1987). The package PAML (Yang 1997) was used for phylogenetic inference involving maximum likelihood methods.

Discussion

The simulation studies performed in this paper present a very competitive phylogenetic reconstruction method based on invariants. If one compares our results on the tree space and those of Huelsenbeck (1995), one sees that the method presented here is highly efficient. Note that this might be a biased comparison as in both papers homogeneous models are used for sequence generation and, although he uses homogeneous methods for inference, our invariant method allows nonhomogeneous models—remember that one generally obtains the best results using the most restricted model that fits the data. There are some limitations of the tree space study performed here, though. For example, this tree space does not consider trees where the inner edge is extremely small or extremely large with respect to the peripheral branches. As Huelsenbeck points out, the usefulness of considering this parameter space can be questioned, but he also gives strong arguments that convinced us to work in this parameter space. Moreover, considering the same parameter space as Huelsenbeck allows one to compare our results with the other methods studied by him.

We would like to comment on the algorithm presented here and, more generally, on all methods based on invariants. First of all, we need to emphasize that the computation of the phylogenetic invariants of a given model just needs to be done once, so the computation need not contribute to the running time of an algorithm using them. Secondly, increasing the size of the sequences does not drop the computational efficiency of the algorithm. Indeed, the sequence’s length only accounts for computing the relative frequencies of the observed patterns (which is something that most algorithms based on evolutionary models must do), but it does not participate in any other part of the algorithm.

A small comment on the election of the 1-norm: we performed simulation studies not presented here to prove that the algorithm performs clearly better with 1-norm than with maximum norm and slightly better than with the Euclidean norm (the “1-norm” of a vector \( x = (x_1, \ldots, x_n) \) is \( \|x\|_1 = \sum_{i=1}^n |x_i| \), the “Euclidean norm” is \( \|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \), and the “maximum norm” is \( \|x\|_\infty = \max_i |x_i| \)). Another consideration that might be important for the computational efficiency of the method is that, in Fourier coordinates, the polynomials considered here are “binomials,” and hence they are easy to evaluate at a given point (so there is no need to worry computationally about the evaluation of the polynomials). Moreover, as it is proved in Sturmfels and Sullivant (2005), these binomials have degree 4 at the most so, again, the computational cost is low. Choosing another generating set of invariants or a different weighting of the polynomials will lead to other results, so this is an issue that should certainly be studied in the future.

We implemented and tested the algorithm presented here only on 4-taxon trees. This seems a limitation of the method, but as the reader may have noticed, the method is universal and could be used to infer the topology of trees with arbitrary number of taxa. However, the computational demands of deducing large phylogenies led us not to develop this algorithm for larger trees. Instead, we suggest that invariants might be a good starting point for quartet methods of phylogenetic inference. In this direction, it is also worth thinking about new tree reconstruction algorithms for arbitrary taxa based on invariants (this is something the authors will surely work on in the future).

In this paper, we focused on the Kimura 3-parameter evolutionary model. However, a full generating set of invariants is known for any group-based model (Sturmfels and Sullivant 2005), a large set of invariants is known for the general Markov model (Allman and Rhodes 2003, 2004a) and for a strand symmetric model (Casanellas and Sullivant 2005), and some invariants are already known for certain rate-class models (Allman and Rhodes 2006). Therefore, the method presented here can be extrapolated to these evolutionary models and in the future further models can be considered. In particular, invariants might also perform well for models allowing site-to-site variation.

As we pointed out in the introduction, invariants-based methods focus on recovering the tree topology and not estimating the parameters. Nevertheless, as Allman and Rhodes (2003, 2004b) say, it is fair to think that phylogenetic invariants may also be useful for parameter recovery.

As we have already mentioned, one of the advantages of the method presented here versus other methods of phylogenetic reconstruction based on evolutionary models is that the model considered here is nonhomogeneous in the sense that the rate matrix is allowed to vary along the different branches of the tree. However, as we are assuming a Kimura model, the base distribution is the same at all nodes of the tree and so the model is stationary (note that invariants methods based on other evolutionary models would not require stationarity of base distribution). For an unrooted tree with \( n \) taxa, our algebraic Kimura model has \( 3(2n - 3) \) free parameters, and it is a special case of the “general Markov model”\(^{12n} \) which involves \( 12(2n - 3) + 3 \) parameters. If one considered a Kimura 3-parameter model (not the algebraic model considered here) allowing different rate matrices along the tree one would have \( 3(2n - 3) + 2n - 4 \) (the extra parameters correspond to time parameters). Other nonhomogeneous models have been considered in the literature, see for example Galtier and Gouy (1998) and Yang and Roberts (1995). The emphasis in these 2 works is put on the nonstationarity hypothesis, and the
maximum likelihood approach taken in most nonhomogeneous methods makes them computationally unfeasible.

**Supplementary Material**

Supplementary figures I–IV are available at *Molecular Biology and Evolution* online (http://www.mbe. oxfordjournals.org/).

**Acknowledgments**

The authors would like to thank B. Sturmfels, L. Pachter, and N. Eriksson for useful comments on a draft version of the paper. We would also like to thank the reviewers for useful comments that improved the paper. Both authors were partially supported by Ministerio de Educación y Ciencia (Ramón y Cajal and Juan de la Cierva programs) and BFM2003-06001. J.F.S. was also supported by MTM2005-01518.

**Literature Cited**


Peter Lockhart, Associate Editor

Accepted October 17, 2006