A new phylogenetic reconstruction method based on invariants

M. Casanellas
J. Fernández-Sánchez
Dpt. Matemàtica Aplicada I
Universitat Politècnica de Catalunya
Av. Diagonal 647. 08028-Barcelona. Spain
marta.casanellas@upc.edu, jesus.fernandez.sanchez@upc.edu

Abstract

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 authors (the initial idea due to Lake [Lake, 1987] and Cavender and Felsenstein [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 non-homogeneous phylogenetic trees.

Introduction

Since the introduction of phylogenetic invariants by Cavender and Felsenstein [Cavender and Felsenstein, 1987] and Lake [Lake, 1987], several attempts to give a generating set of polynomial phylogenetic invariants have been made (see for example [Steel et al., 1993], [Ferreti and Sankoff, 1995]) but it has not been until recently that algebraic geometers have managed to find them all ([Allman and Rhodes, 2004a], [Sturmfels and Sullivant, 2005], [Casanellas and Sullivant, 2005]). Methods based on invariants have already proved to be useful in comparative genomics [Sankoff and Blanchette, 1999]). However, they have not had much success in phylogenetic inference ([Jin and Nei, 1990], [Huelsenbeck, 1995]) because only a small set of invariants has been considered (for example, Lake's method of invariants only used two 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,]). But as Felsenstein explains, invariants are worth more attention for *what they might lead to in the future* ([Felsenstein, 2003]). This future may be the present now since the studies of this paper validate the method based on invariants presented in this report as a promising method. Recently, other methods based on a large set of invariants have also been considered [Eriksson, 2005], [Kim et al.,].

Phylogenetic invariants are equations satisfied by the expected pattern frequencies of a given tree topology T evolving under an evolutionary model. More precisely, if $\mathbf{t}$ is the set of model parameters on T and $p_\alpha(\mathbf{t})$ is the probability of observing the pattern α at the leaves of T, by letting $\mathbf{t}$ vary on an open subset of $\mathbb{R}^d$, the probability vector $p(\mathbf{t}) = (p_{AA\ldots A}(\mathbf{t}), \ldots, p_{TT\ldots T}(\mathbf{t}))$ defines a subset S_T of dimension $\leq d$ of $\mathbb{R}^{4^n}$. A *phylogenetic invariant* is a real-valued continuous $f(x)$ on $\mathbb{R}^{4^n}$ 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 whatever the parameters of the model are, 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 n taxa with enough data, should approximate $p(\mathbf{t})$ for some set of parameters $\mathbf{t}$ 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 that $f(\hat{p})$ is very close to 0. Thus, with probability one, there will be a unique tree topology T for which all its phylogenetic invariants are close to 0 when evaluated at $\hat{p}$. Therefore, using the phylogenetic invariants for tree reconstruction is a consistent method (see [Hagedorn and Landweber, 2000], [Felsenstein, 2003]). A practical introduction to the theory of invariants can be found in the book of J. Felsenstein [Felsenstein, 2003, chapter 22], whereas the book [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 two 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 and its edge lengths and the other is that the evolutionary models underlying the theory of invariants are non-homogeneous. Indeed, it is known that for biological species different rate matrices should be allowed in different lineages. Thus it is essential to have at our disposal phylogenetic methods for reconstructing trees admitting non-homogeneous models (see [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 unrooted trees evolving under the Kimura 3-parameter model of nucleotide substitution [Kimura, 1981]. Actually, we consider an *algebraic* Kimura model: the parameters of the model are the entries of the transition matrices (not of the rate matrices). Hence the model is non-homogenous —because it allows different rate matrices along different lineages— but it is stationary and we always assume that all sites are independent and identically distributed (i.i.d. hypotheses). We performed

simulation studies to prove its efficiency. Our approach to test the efficiency of the method is taken from Huelsenbeck [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 10000.

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

Results

Homogeneous data

The performance of the invariants method studied here on homogeneous trees can be seen in figures 1 and 2. Using the approach of J.P. Huelsenbeck in [Huelsenbeck, 1995] for quartet trees, we considered two branch-length parameters a, b (see figure 6(a)) and simulated data for each couple of parameters. Parameters a and b were varied from 0.01 to 0.75 in increments of 0.02 and 1000 alignments were simulated for each couple (a, b) (see figure 6(b)). The simulated trees evolve under the Kimura 3-parameter model [Kimura, 1981] with a fixed rate matrix of the form

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \cdot & \gamma & \alpha & \beta \\ \gamma & \cdot & \beta & \alpha \\ \alpha & \beta & \cdot & \gamma \\ \beta & \alpha & \gamma & \cdot \end{pmatrix} \end{array}$$

along the tree ($\cdot = -\alpha - \beta - \gamma$). Figures 1 and 2 show the efficiency of the method considered in this paper for two different rate matrices and for nucleotide sequences of lengths 100, 500, 1000 and 10000. The performance of the method is similar in both cases, although the results are slightly better in the first case (notice that in this case the transition:transversion bias is 1:1 against 2.93:1 in figure 2). This difference in the performance is shown, for example, considering the 95% isocline in the graphic corresponding to sequences of length 100.

These figures are to be compared with those shown in Figure A2 of [Huelsenbeck, 1995] (corresponding to phylogenetic inference for sequences generated under a Kimura 2-parameter model of substitution [Kimura, 1980] with 5:1 transition:transversion bias). Though this is of course a biased comparison because our method admits non-homogeneous data and Kimura 3-parameter model, it is worth noticing that 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 only used linear invariants). At least for sequences of length

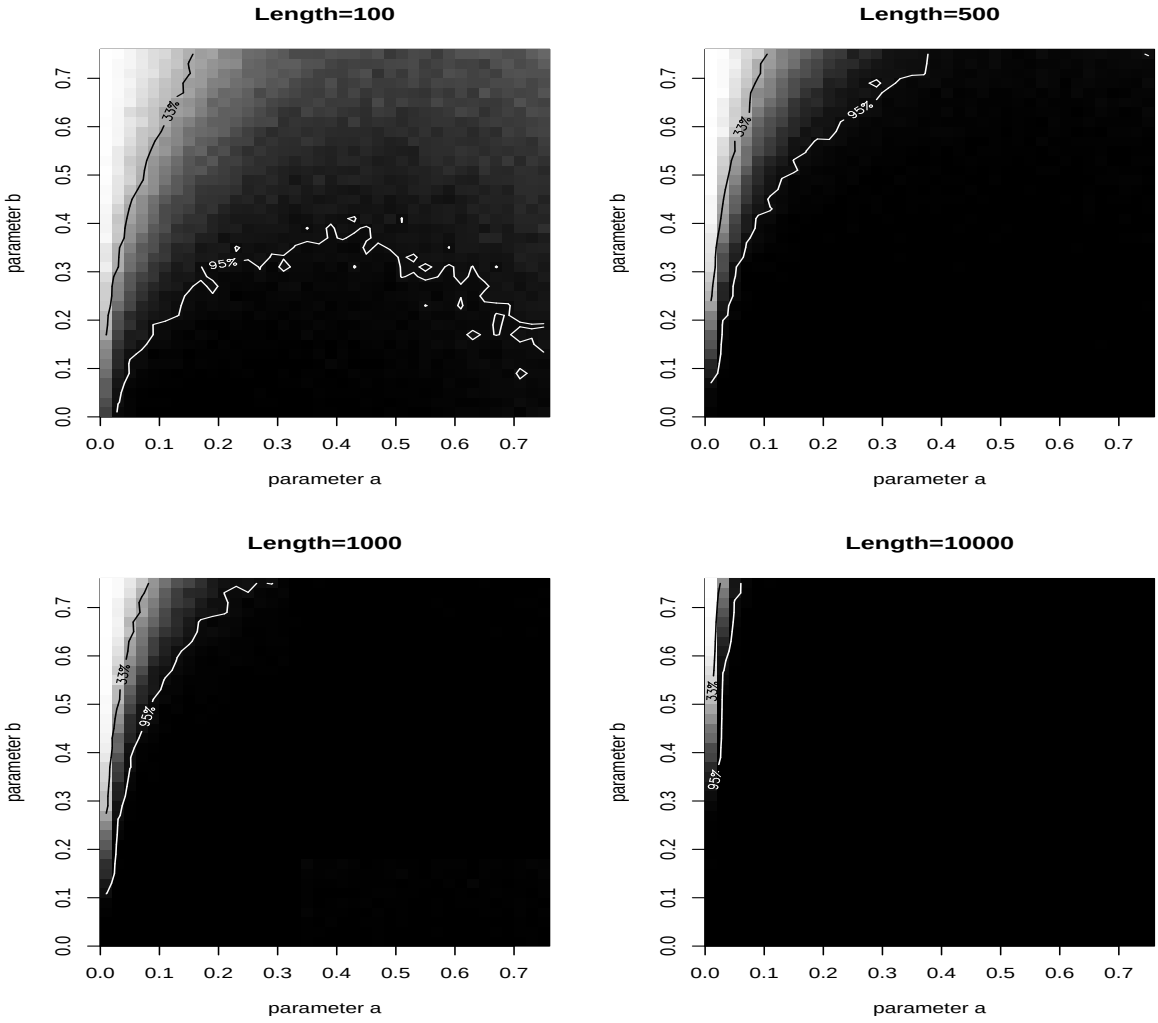

Figure 1: The graphics above represent the probability of reconstructing the correct tree in the parameter space (see Figure 6). Black areas correspond to couples (a, b) for which the tree was correctly estimated, while white regions correspond to couples (a, b) for which the tree is never estimated; grey tones indicate areas of intermediate probability. The 95% isocline is drawn in white, while the 33% is drawn in black. The four graphics above show the results obtained under the Kimura 3-parameter model when using a rate matrix with parameters $\gamma = 0.2, \alpha = 0.5, \beta = 0.3$ (hence a 1:1 transition:transversion bias).

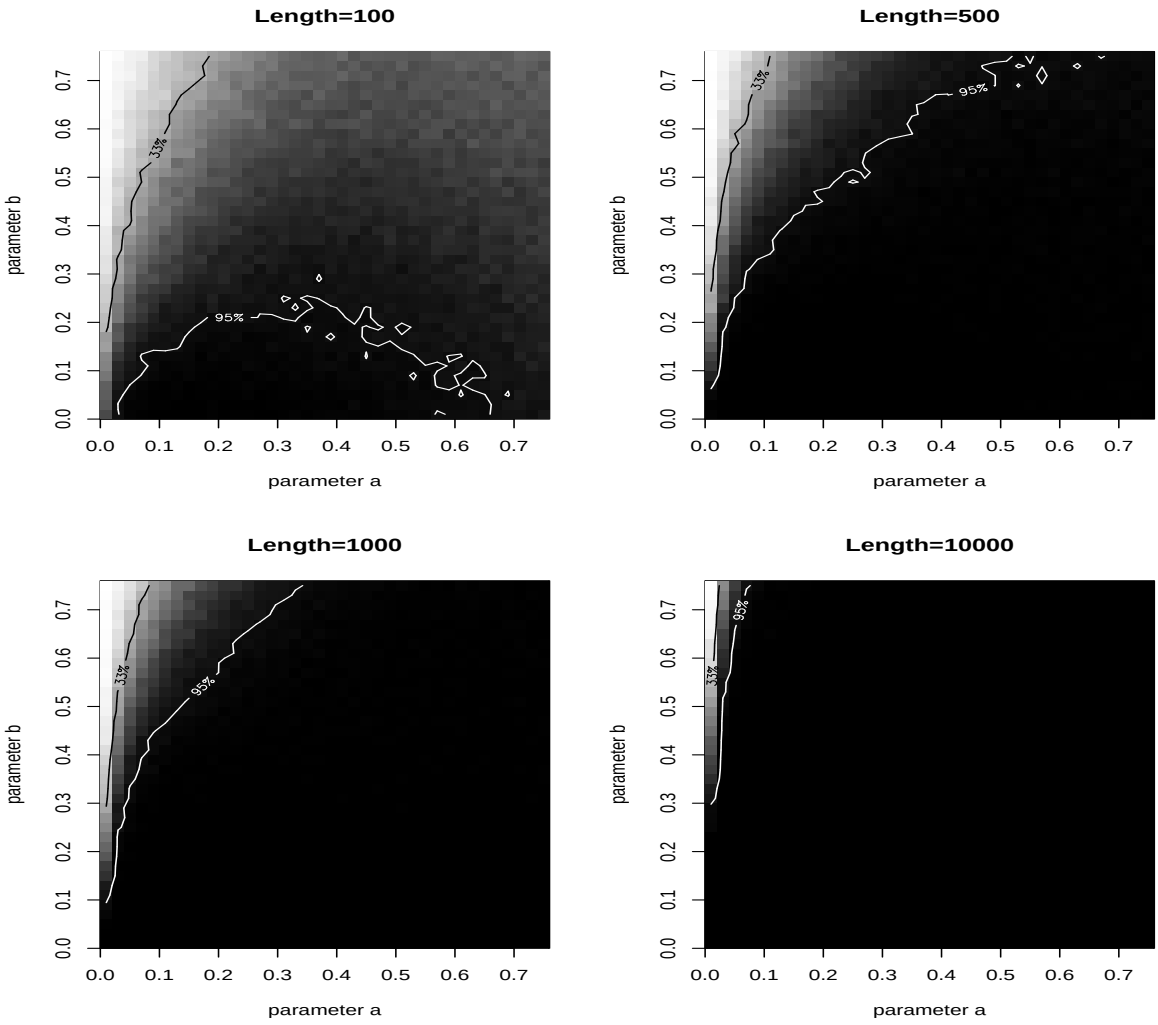

Figure 2: The four graphics above show the results obtained in the four-taxon space when the sequences are generated under the Kimura 3-parameter model of substitution using a rate matrix with parameters $\gamma = 0.12, \alpha = 0.73, \beta = 0.13$ (hence a 2.93:1 transition:transversion bias).

500 or larger, our method performs better than neighbor-joining —referred as Minimum Evolution (Kimura$_1$) in figure A2 of Huelsenbeck's paper. Notice also that the shape of the 95%-isocline for lengths around 500 or larger is quite different from the corresponding shapes in the methods tested by Huelsenbeck: for large values of a (near 0.7), the performance of our method of invariants does not drop drastically (as it happens in most methods considered there).

For length 1000, the efficiency of the invariants method considered here is similar to that obtained for lengths $\geq$ 10000 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 [Hagedorn and Landweber, 2000]).

Non-homogeneous data

We tested the invariants method studied in this paper on non-homogeneous data by comparing it with other methods.

1. Comparison with neigbor-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 3-parameter distance [Kimura, 1981]. As it can be seen in figure 3, considering certain non-homogeneous sets of 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 choosen (see figure 4) 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%. 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.

2. Comparison with maximum likelihood

We compared the invariants method with two versions of maximum-likelihood: the usual maximum likelihood for Kimura 2-parameter [Kimura, 1980] and non-homogeneous maximum likelihood method developped in the package `PAML` [Yang, 1997] for Kimura 2-parameter model (see details in the methods section). 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 4 evolving under the Kimura 2-parameter model. The entries in the rate matrices are functions of a parameter ε that is varied from 1 to 9. When $\varepsilon = 1$ the tree is homogeneous and we studied the efficiency of the three methods as ε increases up to 9.

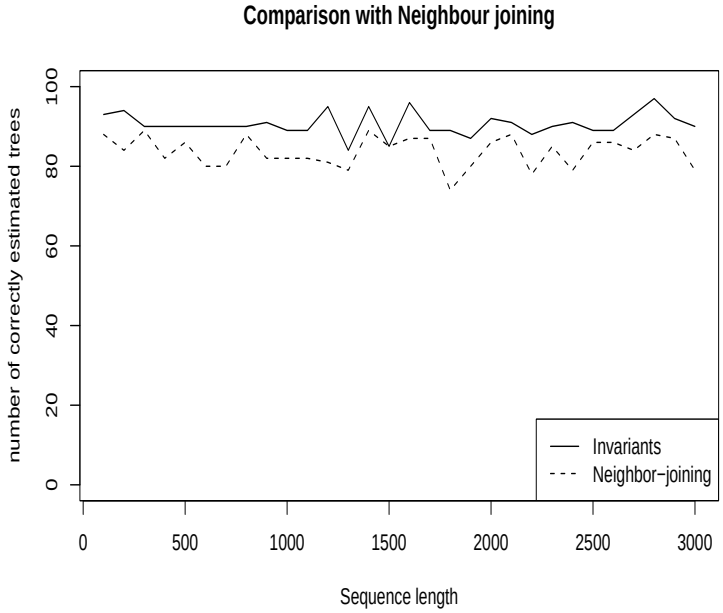

Figure 3: Performance of neighbor-joining and the invariants method on non-homogeneous data for sequence length varying from 100 to 3000. The non-homogeneous tree used for simulations is described in figure 4.

Figure 5 summarizes the results relative to the comparison between our method (`invariants`), the non-homogeneous method in `PAML` for Kimura 2-parameter model (`PAML non-homogeneous`) and the maximum likelihood homogeneous in PAML (`PAML homogeneous`) for Kimura 2-parameter. It shows the percentage of correctly reconstructed trees for each value of parameter ε. As expected, the homogeneous maximum-likelihood algorithm decreases its efficiency as ε increases. Thus, it is not recommendable to use homogeneous methods on data that can be non-homogeneous. Already for $\varepsilon = 5$, the invariants method overtakes the homogenous maximum likelihood algorithm. As it is deduced from figure 5, our method performs worse than the non-homogeneous 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.

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].

We consider the Kimura 3-parameter model [Kimura, 1981] on an unrooted binary (trivalent) tree of n species. The taxa are given by an alignment of n DNA

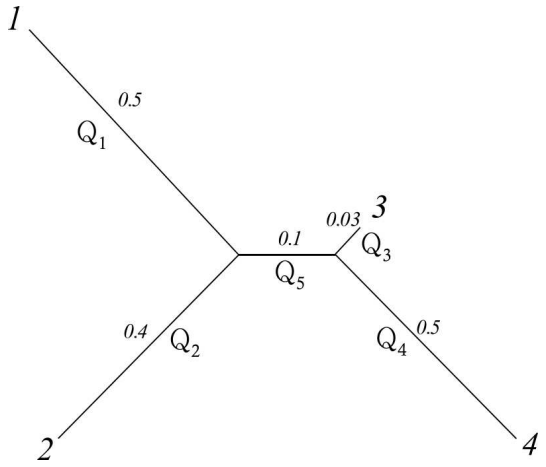

Figure 4: A non-homogeneous tree used for simulations. The numbers labelling each edge correspond to branch lenghts (i.e. expected percent change between the two taxa on the edge). **1.** In comparing our method with neighbor-joining, the parameters $\alpha_i, \beta_i, \gamma_i$ in the Kimura 3-parameter rate matrices Q_i were chosen as: $\gamma_1 = 1, \alpha_1 = 4, \beta_1 = 1$, $\gamma_2 = 5, \alpha_2 = 14, \beta_2 = 3$,$\gamma_3 = 4, \alpha_3 = 15, \beta_3 = 3$, $\gamma_4 = 2, \alpha_4 = 6, \beta_4 = 2$, $\gamma_5 = 2, \alpha_5 = 3, \beta_5 = 1$. **2.** In comparing our method with maximum likelihood algorithm based on the Kimura 2-parameter model, we chose $\gamma = \beta = 1$ in all rate matrices whereas parameter α was set as follows: in Q_1, $\alpha = 4$, in Q_2, Q_4 and Q_5, $\alpha = 3 + \varepsilon^2$, in Q_3, $\alpha = 3 + \varepsilon$. When $\varepsilon = 1$ the tree is homogeneous and we studied the efficiency of three methods as ε increases up to 9.

sequences of length N. A continous time Markov process along the tree is assumed and we consider that all sites are independently and identically distributed (i.i.d. hypotheses). At each branch i the substitution matrix has the form

$$S_i = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} a_i & b_i & c_i & d_i \\ b_i & a_i & d_i & c_i \\ c_i & d_i & a_i & b_i \\ d_i & c_i & b_i & a_i \end{pmatrix} \end{array}$$

for some a_i, b_i, c_i, d_i positive parameters representing the substitution probabilities along the branch i and satisfying $a_i + b_i + c_i + d_i = 1$. Usually a rate matrix Q is fixed and common to the whole tree and S_i is the exponential e^{Qt_i} (for some parameter t_i 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 along different lineages. We want to stress that the parameters of the model we are considering are the entries of the substitution matrices (we should rather speak of an *algebraic* Kimura 3-parameter model, according to the book [Pachter and Sturmfels, 2005, chapters 1 and 4]).

Phylogenetic invariants

Sturmfels and Sullivant [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*. For the Kimura 3-parameter model

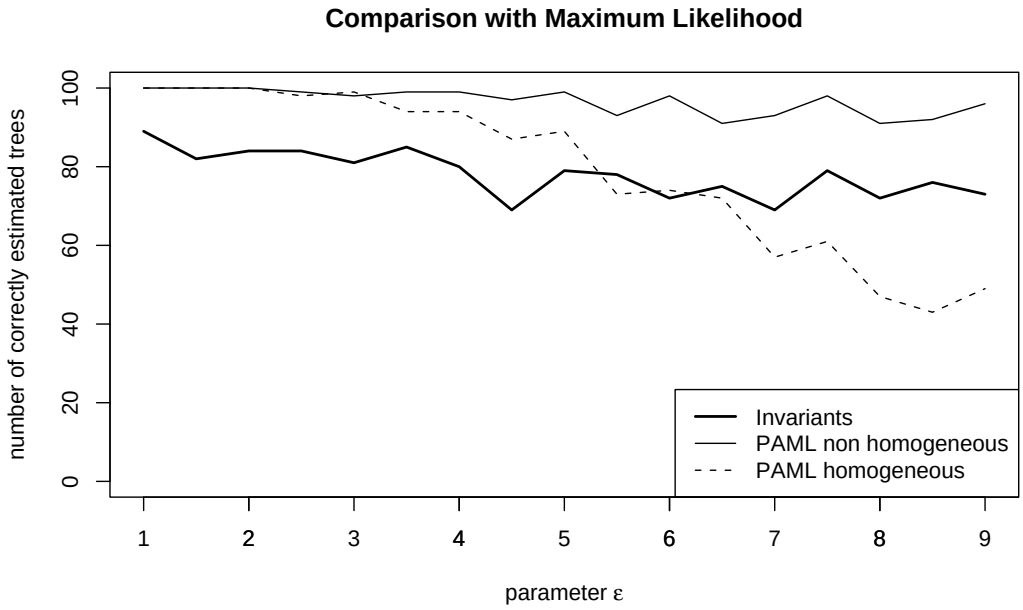

Figure 5: The effect of rate heterogeneity among lineages on method performance. The simulated nucleotide sequences have length 1000 and evolve under the non-homogeneous Kimura 2-parameter model is described in figure 4 (homogeneous for $\varepsilon = 1$). The trees were reconstructed with our method (`invariants`), with `PAML` under a non-homogeneous method (`PAML non-homogeneous`) and the usual `PAML` maximum likelihood homogeneous (`PAML homogeneous`).

on an unrooted 4-taxa tree the ideal of phylogenetic invariants has 8002 minimal generators (see the webpage [Garcia and Porter,], [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 $(p_{A...A}, \ldots, p_{T...T})$ and the phylogenetic invariants are described in terms of the Fourier coordinates. It turns out that the phylogenetic invariants in this case are generated by 144 binomials of degree 2, 1984 binomials of degree 3 and 5874 binomials of degree 4. They can be found in the *small trees* website [Garcia and Porter,] (see [Casanellas et al., 2005] for an explanation on the website), and more precisely at `http://www.math.tamu.edu/~lgp/small-trees/small-trees_30.html`.

Algorithm

Our tree reconstruction algorithm performs the following tasks. Given 4 aligned DNA sequences s_1, s_2, s_3, 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 two 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 these values $\{s_f^T\}_f$, 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 4 sequences of 1000 nucleotides it takes 0.35s on a single 3.0-Ghz processor.

Simulations

To test the sensitivity of our method to branch length variation we considered the approach taken by Huelsenbeck in [Huelsenbeck, 1995], where a large portion of the four-taxon tree space is examined for different phylogenetic reconstruction methods. Therefore, we considered two branch-length parameters and constructed a simulated tree for each couple of parameters. One parameter (parameter a) assigns the branch length to the internal branch and two opposite peripheral branches, and the other parameter (parameter b) assigns the branch length to the two remaining branches (see figure 6).The parameters a and b were varied from 0.01 to 0.75 in increments of 0.02, so that the parameter space studied here includes 1296 different combinations of branch lengths (figure 6).

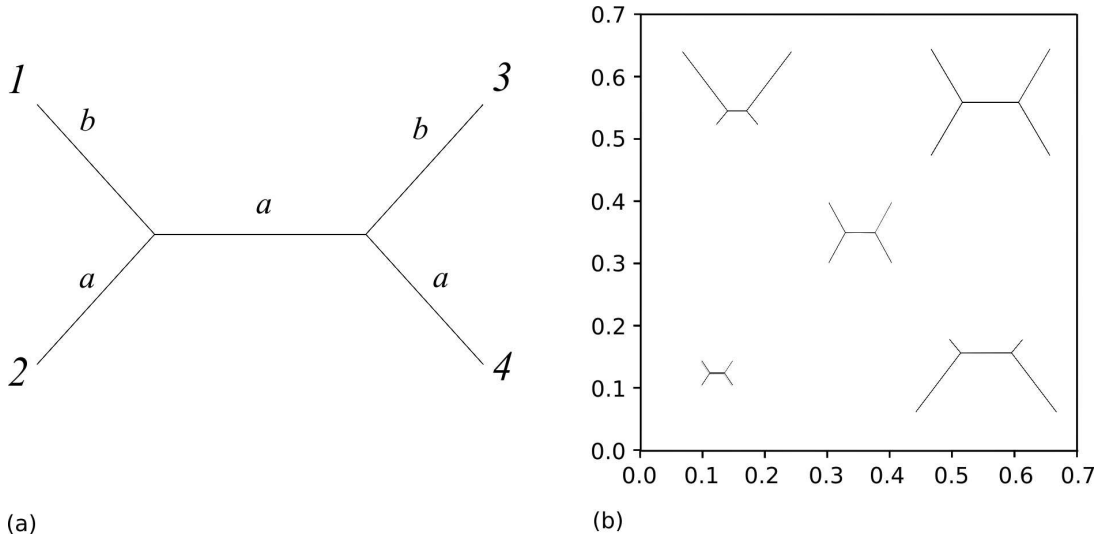

Figure 6: (a) The parameter space for the unrooted four-taxon tree. Parameter a is plotted in the abscissa, while parameter b in the ordinate. Changes in the length of the branches a and b are varied from 1% to 75% in increments of 2%. (b) Unrooted 4-taxa tree. The branches labelled with the same letter (either a or b) are supposed to have the same expected percent of changes. The taxa are named with arabic numbers, 1 to 4.

The simulations of this study were obtained using the program Seq-Gen v1.3.2 [Rambaut and Grassly, 1997]. We performed two different sets of simulations in the parameter space. In the first set we consider the following calibrated rate matrix for each branch:

$$\begin{pmatrix} \cdot & 0.2 & 0.5 & 0.3 \\ 0.2 & \cdot & 0.3 & 0.5 \\ 0.5 & 0.3 & \cdot & 0.2 \\ 0.3 & 0.5 & 0.2 & \cdot \end{pmatrix}$$

where $\cdot = -0.2 - 0.5 - 0.3$. For each assignment of branch lengths in the tree space we simulated 1000 alignments for which we used our algorithm to reconstruct the tree, so that in total 1296000 tests of our algorithm have been performed. The results corresponding to this simulation appear in figure 1.

In the second set of simulations we considered the following rate matrix along all branches

$$\begin{pmatrix} \cdot & 0.12 & 0.73 & 0.13 \\ 0.12 & \cdot & 0.13 & 0.73 \\ 0.73 & 0.13 & \cdot & 0.12 \\ 0.13 & 0.73 & 0.12 & \cdot \end{pmatrix}$$

where $\cdot = -0.12 - 0.73 - 0.13$. This rate matrix was obtained from [Al-Aidroos and Snir, 2005] and it corresponds to the maximum likelihood estimate from an alignment of homologous sequences of eight vertebrates under a Kimura 3-parameter model. Again, for each assignment of branch lengths in the tree space we did 1000 tests of our algorithm. The results of this set of simulations are presented in figure 2.

As we mentioned above, the hypotheses of the model allow for different rate matrices along different lineages in the tree. In other words, our method admits a non-homogeneous model. 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). To test the non-homogeneity hypothesis in our phylogenetic reconstruction method we performed the following study.

1. Comparison with neigbor-joining

We considered the tree in figure 4, with rate matrices as given in figure caption. For each lenght from 100 to 3000, in intervals of 100, we generated 100 sets of data according to the model specified in figure 4. The simulations were made using Seq-Gen, as this software allows to keep record of ancestral sequences and it is possible to generate trees with a given ancestor sequence at the root. The we reconstructed the tree using the invariants method and neighor-joining (see results in figure 3). We used the algorithms implemented in the package `APE` [Paradis et al., 2006] v1.8-3 of `R` [Team, 2005] v2.1.1 to compute Kimura 3-parameter distance [Kimura, 1981] and perform neighbor-joining algorithm [Saitou and Nei, 1987].

2. Comparison with maximum likelihood

Finally, to perform the comparison of our program against a maximum likelihood we considered the tree in figure 4 (branch lengths represent expected percentage of substitutions per site and were chosen so that none of the methods considered below performed perfectly). The rate matrices in figure 4 correspond to Kimura 2-parameter model and are

$$Q_1 = \begin{pmatrix} \cdot & 1 & 4 & 1 \\ 1 & \cdot & 1 & 4 \\ 4 & 1 & \cdot & 1 \\ 1 & 4 & 1 & \cdot \end{pmatrix},$$

$$Q_2 = Q_4 = Q_5 = \begin{pmatrix} \cdot & 1 & 3+\varepsilon^2 & 1 \\ 1 & \cdot & 1 & 3+\varepsilon^2 \\ 3+\varepsilon^2 & 1 & \cdot & 1 \\ 1 & 3+\varepsilon^2 & 1 & \cdot \end{pmatrix},$$

$$Q_3 = \begin{pmatrix} \cdot & 1 & 3+\varepsilon & 1 \\ 1 & \cdot & 1 & 3+\varepsilon \\ 3+\varepsilon & 1 & \cdot & 1 \\ 1 & 3+\varepsilon & 1 & \cdot \end{pmatrix}.$$

We let parameter ε vary from 1 to 9.0 in intervals of 0.5. For each value of ε Seq-Gen generated 100 replicates of data. We reconstructed the tree with our algorithm and with two algorithms in PAML [Yang, 1997]: maximum-likelihood under a Kimura 2-parameter model homogeneous (option `nhomo=0` in baseml control file) and the non-homogeneous maximum likelihood Kimura 2-parameter (option `nhomo=0` in baseml control file). Option *nhomo=2* in PAML allows for different transition/transversion ratio for different branches in the tree. Notice that this option in PAML has no effect on Kimura 3-parameter model, so we had to simulate data under a Kimura 2-parameter model. The results of this test can be seen in figure 5.

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 [Huelsenbeck, 1995] (though it is a biased comparison as he uses Jukes-Cantor or Kimura 2-parameter models for sequence generation and our model is non-homogeneous), one sees that the method presented here is highly efficient. 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 to the other methods studied by him.

We would like to make some considerations on the algorithm presented here and, more generally, on all methods based on invariants. First of all we need to emphasize that the phylogenetic invariants of a given model just need to be computed once in your life. Then one can use them for phylogenetic inference as we have done in this paper. Secondly, increasing the size of the sequences does not drop the computational efficiency of the algorithm. Indeed, the sequences 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 the 1-norm than with the maximum norm (which takes into account only the distance to one of the hypersurfaces containing the variety), and slightly better than with the euclidean norm. 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 most so, again, the computational cost is low.

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]) and a large set of invariants is known for the general Markov model ([Allman and Rhodes, 2003], [Allman and Rhodes, 2004a]). Therefore, the method presented here can be extrapolated to these evolutionary models and in the future further models can be considered.

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 say ([Allman and Rhodes, 2003], [Allman and Rhodes, 2004b]), 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 non-homogeneous in the sense that the

rate matrix is allowed to vary along the different branches of the tree. However, the base distribution is the same at all nodes of the tree and so the model is stationary. 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* 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-3 (the extra parameters correspond to time parameters). Other non-homogeneous models have been considered in the literature, see for example [Galtier and Gouy, 1998] and [Yang and Roberts, 1995]. The emphasis in these two papers is put on the non-stationarity hypothesis, and the maximum likelihood approach taken in most non-homogeneous methods makes them computationally unfeasible.

Acknowledgments: The authors would like to thank B. Sturmfels, L. Pachter and N. Eriksson 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, respectively), and BFM2003-06001. Second author was also supported by MTM2005-01518.

References

[Al-Aidroos and Snir, 2005] Al-Aidroos, J. and Snir, S. (2005). Analysis of point mutations in vertebrate genomes. In Pachter, L. and Sturmfels, B., editors, *Algebraic Statistics for computational biology*, chapter 21, pages 375–386. Cambridge University Press.

[Allman and Rhodes, 2003] Allman, E. and Rhodes, J. (2003). Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186(2):113–144.

[Allman and Rhodes, 2004a] Allman, E. and Rhodes, J. (2004a). Phylogenetic ideals and varieties for the general Markov model. Preprint, http://arxiv.org/abs/math.AG/0410604.

[Allman and Rhodes, 2004b] Allman, E. and Rhodes, J. (2004b). Quartets and parameter recovery for the general Markov model of sequence mutation. *AMRX Applied Mathematics Research Express*, 2004(4):107–131.

[Casanellas et al., 2005] Casanellas, M., Garcia, L., and Sullivant, S. (2005). Catalog of small trees. In Pachter, L. and Sturmfels, B., editors, *Algebraic Statistics for computational biology*, chapter 15. Cambridge University Press.

[Casanellas and Sullivant, 2005] Casanellas, M. and Sullivant, S. (2005). The strand symmetric model. In Pachter, L. and Sturmfels, B., editors, *Algebraic Statistics for computational biology*, chapter 16. Cambridge University Press.

[Cavender and Felsenstein, 1987] Cavender, J. and Felsenstein, J. (1987). Invariants of phylogenies in a simple case with discrete states. *J. Classification*, 4:57–71.

[Eriksson, 2005] Eriksson, N. (2005). Tree construction using singular value decomposition. In Pachter, L. and Sturmfels, B., editors, *Algebraic Statistics for computational biology*, chapter 19, pages 347–358. Cambridge University Press.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.

[Felsenstein, 2003] Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates, Inc.

[Ferreti and Sankoff, 1995] Ferreti, V. and Sankoff, D. (1995). Phylogenetic invariants for more general evolutionary models. *J. theor. Biol.*, 147-162.

[Galtier and Gouy, 1998] Galtier, N. and Gouy, M. (1998). Inferring pattern and process: maximum likelihood implementation of a non-homogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.*, 154(4):871–879.

[Garcia and Porter,] Garcia, L. and Porter, J. Small phylogenetic trees webpage. http://bio.math.berkeley.edu/ascb/chapter15/.

[Hagedorn and Landweber, 2000] Hagedorn, T. and Landweber, L. (2000). Phylogenetic invariants and geometry. *J. theor. Biol.*, 205:365–376.

[Huelsenbeck, 1995] Huelsenbeck, J. (1995). Performance of phylogentic methods in simulation. *Syst. Biol.*, 44:17–48.

[Jin and Nei, 1990] Jin, L. and Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.*, 7:82–102.

[Kim et al.,] Kim, Y. R., Kwon, O., Paeng, S., and Park, C. Computational methods in constructing phylogenetic trees using SVDs of flattenings. Preprint July 2006.

[Kimura, 1980] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120.

[Kimura, 1981] Kimura, M. (1981). Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Nat. Acad. Sci. , USA*, 78:454–458.

[Lake, 1987] Lake, J. (1987). A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony. *Mol. Biol. Evol.*, 4:167–191.

[Pachter and Sturmfels, 2005] Pachter, L. and Sturmfels, B., editors (2005). *Algebraic Statistics for computational biology.* Cambride University Press. ISBN 0-521-85700-7.

[Paradis et al., 2006] Paradis, E., Strimmer, K., Claude, J., Jobb, G., Opgen-Rhein, R., Dutheil, J., Noel, Y., Bolker, B., and Lemon, J. (2006). *ape: Analyses of Phylogenetics and Evolution.* R package version 1.8-3.

[Rambaut and Grassly, 1997] Rambaut, A. and Grassly, N. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13:235–238.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425.

[Sankoff and Blanchette, 1999] Sankoff, D. and Blanchette, M. (1999). Phylogenetic invariants for genome rearrangements. *J. Comput. Biol.*, 6:431–445.

[Steel et al., 1993] Steel, M., Székely, L., Erdős, P., and Waddell, P. (1993). A complete family of phylogenetic invariants for any number of taxa under kimura's 3st model. *NZ J. Bot.*, 31:289–296.

[Sturmfels and Sullivant, 2005] Sturmfels, B. and Sullivant, S. (2005). Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12:204–228.

[Team, 2005] Team, R. D. C. (2005). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Yang, 1997] Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *CABIOS*, 15:555–556.

[Yang and Roberts, 1995] Yang, Z. and Roberts, D. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.*, 12:451–458.

[Yang and Yoder, 1999] Yang, Z. and Yoder, A. D. (1999). Estimation of the transition/transversion rate bias and species sampling. *J. Mor. Evol.*, 48:274–283.