

# Simulation studies of algebraic invariants for phylogeny

Nicholas Eriksson

May 13, 2004

## 1 Introduction

The reconstruction of the evolutionary history of a group of taxa is a major problem in computational biology. For example, given (some portion) of the DNA sequence for several taxa, we wish to construct how this sequence has evolved. Construction of this phylogenetic tree is extremely useful for deducing relationships (e.g., biochemical or structural) among the taxa. Despite much progress in recent years, it is still an important problem to develop methods which are fast, accurate, and do not depend on untestable assumptions.

Methods such as Neighbor Joining involve the calculation of a distance matrix from the input sequences, which relies on assumptions about the stochastic model of evolution. However, Neighbor Joining is both fast and accurate in practice, so it is very popular.

In this paper, we investigate another method for deducing phylogenetic relationships which uses tools from algebraic geometry and relies on no assumptions other than evolution following a certain family of Markov models.

First, we review the commonly used Markov models in Section 2. In Section 3, we introduce algebraic invariants, show how some of them can be effectively computed, and use them for an example with 7 yeast species. Section 4 deals with the design of our simulation study. Results of the simulations are given in Section 5. We end with conclusions and open problems in Section 6.

## 2 Markov models of evolution

In this section, we present the model of evolution that we assume (the general Markov model). Let  $T$  be a rooted tree with  $n$  leaves. To each node  $v$  of  $T$ , we associate a  $k$ -ary random variable  $X_v$ . In this paper,  $k$  will be 4 (A,C,G,T), but other common values are 2 (the purines and pyrimidines) or 20 (the amino acids). The interior nodes of the tree are hidden random variables, the leaves are observed. There is a distribution  $\pi$  on the root node which we assume to be uniform. We label each edge of the tree with a length. Then the random

variables are updated by a continuous time Markov process according to transition matrices  $A^e$  chosen from some family on each edge. For the purposes of this paper, we assume nothing about these matrices. However, many specialized families have been proposed, several of which will be discussed in Section 3.

Given the joint probabilities on the observed random variables (the leaves), we wish to infer the structure of the tree. We write  $p_{i_1 \dots i_n} := P(X_1 = i_1, \dots, X_n = i_n)$  as these joint probabilities.

### 3 Algebraic invariants

It was originally proposed by Cavendar and Felsenstein [4] that the algebraic relations between the joint probabilities could be used to reconstruct the evolutionary tree. For any tree based model, there exists a set of algebraic relations between the joint probabilities that uniquely characterizes the model.

Much work has gone into finding partial or complete lists of invariants for various models, see [5, 6, 7, 12, 15]. Recently, algebraists have become interested in the problem, leading to results for the Jukes-Cantor, Kimura, and general Markov models [13, 1, 2]. Briefly, the specialized models such as Jukes-Cantor assume some special form for the transition matrices, e.g., that all transversions/transitions have the same probability.

**Definition 1.** The *algebraic invariants* of a tree  $T$  and a Markov model  $\mathcal{M}$  are the algebraic relations between the joint probabilities of the observations at the leaves of  $T$ . These relations form an ideal in the polynomial ring generated by the joint probabilities.

Algebraic invariants possess several impediments to actual use for phylogeny. First of all, only recently have all the invariants been found for common models. Second, there are many invariants, and it is not clear how to evaluate all of them. There are exponentially many invariants in exponentially many unknowns (the joint probabilities), and exponentially many trees to check. Due to these problems, invariants have not been used in practice and little work has been done on their practicality.

However, there is a determinantal representation of the ideal of invariants for the binary general Markov model which provides an effective condition to test invariants.

To introduce this representation, we need to introduce the idea of a “split”. Given a tree  $T$ , every edge induces a *split* of the set of leaves corresponding to the two connected components of the tree obtained by removing that edge.

For every split  $(\mathcal{A}, \mathcal{B})$  of the leaves of the tree, we associate a  $4^{|\mathcal{A}|} \times 4^{|\mathcal{B}|}$  matrix as follows. The rows of the matrix are indexed by the functions  $\mathcal{A} \rightarrow \{A, C, G, T\}$  and the columns by the functions  $\mathcal{B} \rightarrow \{A, C, G, T\}$ . The entries of the matrix are the joint probabilities of observing the corresponding bases at  $\mathcal{A}$  and  $\mathcal{B}$ .

Then it is conjectured (see [9]) that the algebraic invariants for the general Markov model on this tree are exactly the 5 by 5 determinants of the matrices

corresponding to all splits in the tree. Recall that the 5 by 5 determinants of a matrix vanish precisely when the matrix has rank 4.

A proof of this conjecture for binary data (where the 3 by 3 determinates vanish) has been announced by Allman and Rhodes, see [1, 2] for work leading up to this.

**Example 2 (A split on 4 taxa).** The split  $\{1, 3\}$  and  $\{2, 4\}$  generates the  $16 \times 16$  matrix where the rows are indexed by bases of taxa 1 and 3 and the columns by bases of taxa 2 and 4.

	AA	AC	AG	AT	CA	CC	...
AA	$p_{AAAA}$	$p_{AAAAC}$	$p_{AAAAG}$	$p_{AAAAT}$	$p_{ACAA}$	$p_{ACAC}$	...
AC	$p_{AACA}$	$p_{AACC}$	$p_{AACG}$	$p_{AACT}$	$p_{ACCA}$	$p_{ACCC}$	...
AG	$p_{AAGA}$	$p_{AAGC}$	$p_{AAGG}$	$p_{AAGT}$	$p_{ACGA}$	$p_{ACGC}$	...
AT	$p_{AATA}$	$p_{AATC}$	$p_{AATG}$	$p_{AATT}$	$p_{ACTA}$	$p_{ACTC}$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

The according to the above conjecture, the split  $\{1, 3\}, \{2, 4\}$  occurs in the phylogenetic tree for these taxa exactly when this matrix has rank 4.

Given a sample from the distribution  $p_{i_1 \dots i_n}$ , we can test how close the matrices corresponding to splits are to rank 4 by computing the singular value decomposition.

**Definition 3.** The *singular value decomposition* of a  $n$  by  $m$  matrix  $A$  ( $m \geq n$ ) is given by

$$A = U\Sigma V^T$$

where  $U$  is a  $m$  by  $m$  matrix and  $V$  is a  $n$  by  $n$  matrix, both of which have orthogonal columns, and  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$  is an  $m$  by  $n$  matrix with diagonal elements the *singular values* of  $A$ .

The singular value decomposition has the property that  $\sigma_m$  is the distance in the matrix 2-norm to the closest rank  $m - 1$  matrix.

**Example 4.** Starting with a multiple alignment from `mavid` [3] of the entire DNA sequence from 7 species of yeast we built the table of joint probabilities (ignoring gaps). In this table, 12013 of the  $4^7 = 16384$  entries were non-zero. Each sequence was about 15 mbp long, after ignoring gaps there were 1107225 bases in each sequence. The correct tree for these species shown in Figure 1. We calculated the matrices for all possible splits of the data into sets of size 3 or 4. For each, we calculated the singular value decomposition and used the fifth singular value as the “score” of the split. A smaller score corresponds to a better split. Using invariants, the split by the dotted line scored best (at 12,176), among all splits, followed by the split which switches kudriavzevii and mikatae (score 16,580). This is consistent with accepted wisdom.

## 4 Experimental design

The main result of this project is simulation studies for invariants. First, we constructed a tree with 10 leaves, see Figure 2. The edge lengths were chosen

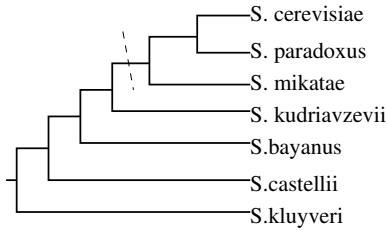


Figure 1: The correct phylogenetic tree for the yeast species and the split identified by invariants as the best.

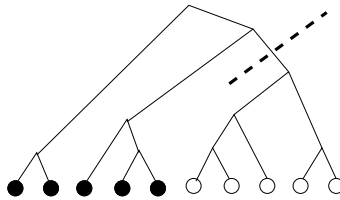


Figure 2: The tree with 10 leaves we tested, and a split of this tree into two sets of size 5.

to be normally distributed with mean  $\bar{\lambda}_e = 0.25, 0.05, 0.005$ . The edge length corresponds to the expected number of mutations per character at an edge.

The program `seq-gen` [10] was used to simulate data from the reversible Markov model. We simulated 10 data sets for sequence lengths ranging from 50 to 8000 for each evolution rate. The evolution rates and sequence lengths were chosen similarly to [11], which tested various methods of phylogenetic tree reconstruction.

For each data set, the SVD was computed for every split of the 10 taxa into two sets of size 5. That is, for each of the  $\binom{10}{5} = 252$  such splits, a 1024 by 1024 matrix was computed from the data. This matrix was output in sparse format. Then a fortran program using the `PROPACK` library was used to compute the singular value decomposition. The sum of the first 100 singular values minus the sum of the first 4 singular values was taken as the “score” of the split. If the split that actually occurred in the tree had the best score out of all 252 splits, the trial was deemed a success.

## 5 Experimental results

Recall that  $\bar{\lambda}_e$  denotes the expected number of events on a random edge in the model tree.

At a high evolution rate ( $\bar{\lambda}_e = 0.25$ ), a sequence of length 5000 was needed before we could identify the split, see Figure 3. This is a quite long sequence,

as most tree reconstruction methods need sequences of length at most 1000 to perform well. The standard deviation of the scores was quite low, around 1–3% for sequences of size over 1000.

At a lower evolution rate ( $\bar{\lambda}_e = 0.05$ ), a sequence of length only 100 was needed to reliably identify the split, see Figure 3. The standard deviation was again low (around 3–5%) for reasonably long sequences.

At the lowest evolution rate ( $\bar{\lambda}_e = 0.005$ ), a sequence of length 500-2000 was needed to identify the split, see Figure 3. However, it seemed there were numerical problems with the SVD for this test, as the standard deviation in the scores was over 50%.

In comparison, with a sequence length of 500, Neighbor Joining constructed the tree right 95% of the time for  $\bar{\lambda}_e = 0.25$ , 98% of the time for  $\bar{\lambda}_e = 0.05$ , and 68% of the time for  $\bar{\lambda}_e = 0.005$ . Thus our results are at least roughly comparable to neighbor joining (although it should be noted that constructing the right tree is a much harder task than finding one split of the data).

## 6 Conclusions and open problems

It remains to be seen if phylogenetic invariants can actually be used to reconstruct phylogenetic trees. However, we have shown that they can possibly be useful in checking splits in unresolved trees. Algebraic invariants produce the correct split for sequences of relatively small length at reasonable rates of evolution. They perform better with lower rates of evolution, although there seem to be numerical problems with extremely low rates. Invariants also fare well in practice, producing the correct split for the 7 yeast species.

Furthermore, we have shown that algebraic invariants which are given by rank conditions can be effectively evaluated for 10 taxa using numerical linear algebra techniques. However, the size of the matrices grows exponentially in the number of taxa. It is not clear how large of an example we can do, but these methods should extend to 15-20 taxa. It is an important practical question whether we can use binary sequences (i.e., only keeping track of purine/pyrimidine) instead of DNA sequences. If this is possible, algebraic invariants may be useful for up to 40 taxa.

Further work will involve testing the invariants of [13] for the Jukes-Cantor and Kimura models.

Finally, an important open problem is whether trees can be reconstructed from knowledge of their splits. Since there are  $2^{n-1} - 1$  non-trivial splits of  $n$  taxa, we can't hope to test all splits for  $n$  over about 10. However, possibly partial knowledge of good splits can be used to reconstruct phylogenetic trees.

## Acknowledgements

Thanks to Jim Demmel for help with computing the singular value decomposition and Seth Sullivant for useful discussions.

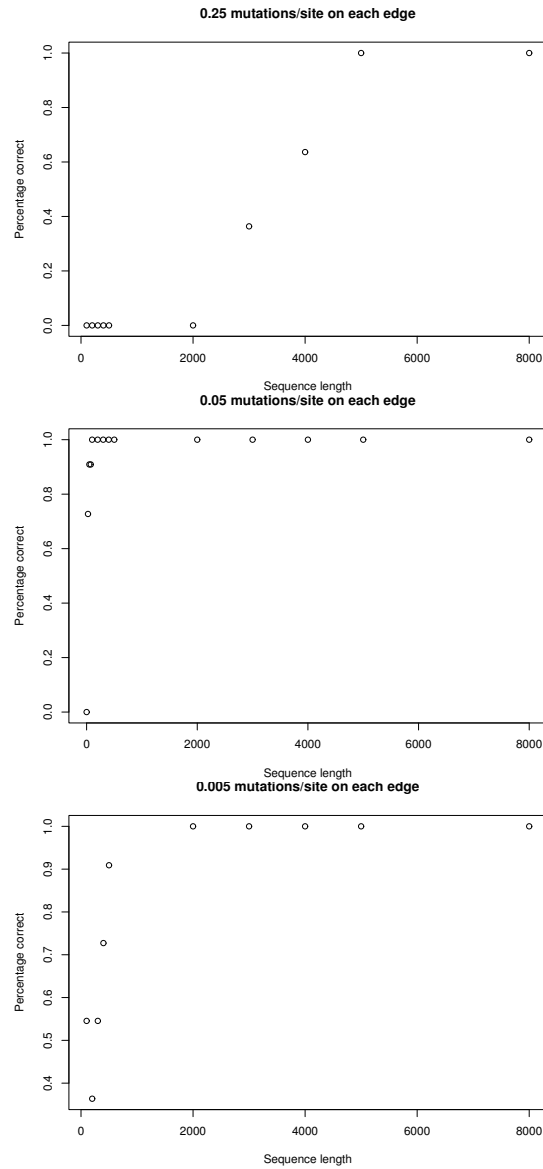


Figure 3: Correct identifications of the split at various sequence lengths for a high, medium, and low rate of evolution.

## References

- [1] E. Allman and J. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences* **186** (2003) 133-144,
- [2] E. Allman and J. Rhodes. Quartets and parameter recovery for the general Markov model of sequence mutation, preprint, 2003.
- [3] Bray, N. and Pachter, L., MAVID: Constrained ancestral alignment of multiple sequences, *Genome Research*, **14**:693-699 (2004).
- [4] J. A. Cavender and J. Felsenstein. Invariants of phylogenies: a simple case with discrete states. *Journal of Classification* **4** (1987) 57-71.
- [5] S. Evans and T. Speed. Invariants of some probability models used in phylogenetic inference. *Annals of Statistics* **21** (1993) 355-377.
- [6] V. Ferretti and D. Sankoff. Phylogenetic invariants for more general evolutionary models, *Journal of Theoretical Biology* **173** (1995) 147-162.
- [7] T. Hagedorn. Determining the number and structure of phylogenetic invariants. *Advances in Applied Mathematics* **24** (2000) 1-21.
- [8] J. A. Lake. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution* **4** (1987) 167-191.
- [9] L. Pachter and B. Sturmfels. The tropical geometry of statistical models, [q-bio.QM/0311009](https://arxiv.org/abs/q-bio.QM/0311009).
- [10] Rambaut, A. and Grassly, N. C. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* (1996)
- [11] K. St. John, T. Warnow, B Moret, and L. Vawter, Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor joining. *Journal of Algorithms* **48** (2003) 173-193.
- [12] D. Sankoff and M. Blanchette. Phylogenetic invariants for genome rearrangements. *Journal of Computational Biology* **6** (1999) 431-445.
- [13] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. [q-bio.PE/0402015](https://arxiv.org/abs/q-bio.PE/0402015)
- [14] M. Steel and Y. Fu. Classifying and counting linear phylogenetic invariants for the Jukes Cantor model. *Journal of Computational Biology* **2** (1995) 39-47.
- [15] M. Steel, L. Székely, P. Erdős, and P. Waddell. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *NZ Journal of Botany* **13** (1993) 289-296.