

Tree construction using singular value decomposition

Nicholas Eriksson ¹

Keywords: phylogenetic trees, Markov models, phylogenetic invariants.

1 Introduction.

We present a new, statistically consistent algorithm for phylogenetic tree construction that uses the algebraic theory of statistical models (as developed in [5]). Our basic tool is the *Singular Value Decomposition* (SVD) from numerical linear algebra (see [3]). Starting with a multiple alignment of n species, we show that the SVD allows us to decide whether a split of the species occurs in their phylogenetic tree. Using this fact, we have developed an algorithm (jointly with Sagi Snir) to construct a phylogenetic tree by computing only n^2 SVD's.

Our algorithm only assumes that evolution follows a Markov model on a binary tree and that evolutions happens independently at different sites of the genome. No assumptions are made about the shape of the transition matrices.

The algorithm uses *phylogenetic invariants*. Such polynomials have been studied for years (e.g., [1, 2]) and have been used to infer phylogenies on four and five taxa ([7]), but have been widely considered impractical. However, our algorithm is very fast in practice on trees with up to 15-25 taxa. It shows promise for real data because it does not assume the existence of a global rate matrix; for example, it places the rodents correctly more often than other methods do. We have implemented this algorithm using the SVDLIBC library ([6]) and have done extensive testing with simulated and real data.

2 Tree construction algorithm.

Fix a tree with n leaves. At each node of the tree there is a random variable with m states (usually $m = 4$, $\{\text{A, C, G, T}\}$). The leaves are observed, the interior nodes are hidden. We will write the joint probabilities of an observation on the leaves as $p_{i_1 \dots i_n}$. That is, $p_{i_1 \dots i_n}$ is the probability that leaf j is observed to be in state i_j .

A split A, B in a tree is a partition of the leaves obtained by removing an edge of the tree. A *flattening* along a split A, B is the $m^{|A|}$ by $m^{|B|}$ matrix where the rows are indexed by the possible states for the leaves in A and the columns are indexed by the possible states for the leaves in B . The entries of this matrix are given by the joint probabilities of observing the given pattern at the leaves.

Theorem 1. *Flattenings along splits in the tree have rank m , while flattenings along partitions that are not splits have higher rank.*

Algorithm 2 (Tree construction with SVD, joint with S. Snir).

Input: A multiple alignment of genome data from n species, from an alphabet Σ with m states.

Output: A phylogenetic tree.

¹University of California, Berkeley, California. E-mail: eriksson@math.berkeley.edu

Initialization: Compute joint probabilities $p_{i_1 \dots i_n}$. That is, count occurrences of each possible column of the alignment, ignoring columns with characters not in Σ . Store the results in a sparse format.

Loop: For k from n down to 2.

For each of the $\binom{k}{2}$ pairs of species compute the SVD for the flattening along the partition {pair}, {other $k - 2$ species}. Pick the pair that is closest to rank m (according to the Frobenius norm) and join this pair together in the tree. That is, consider this pair as a single element when picking pairs at the next step.

Theorem 3. *Algorithm 2 is statistically consistent. That is, as the probability distribution converges to a distribution that comes from the general Markov model on a binary tree T , the probability that it outputs T goes to 1.*

3 Performance studies.

We simulated data of various lengths under the general reversible model for various trees with up to 20 taxa with various branch lengths. We ran all tests using the SVD algorithm as well as two algorithms from the PHYLIP ([4]) package: neighbor joining and a maximum likelihood algorithm (`dnaml`). Performance was comparable, although our SVD algorithm was generally less successful in reconstructing trees. However, the SVD algorithm actually performed significantly better on binary data than on DNA data; the other algorithms performed worse on binary data.

Tree construction methods that use genomic data usually misplace the rodents on the tree of life. The reasons for this are not entirely known, but it could be because tree construction methods generally assume the existence of a global rate matrix for all the species, however, rat and mouse have mutated faster than the other species [8].

In contrast, our method does not assume anything about the rate matrix and performs better on real data sets than `dnaml`. While it did not construct the correct tree a majority of the time, it came much closer on average than `dnaml`, which almost never constructed the correct tree. We believe that this new algorithm will be promising for situations with binary data and situations where additional assumptions beyond the Markov process of evolution at independent sites are not warranted.

References

- [1] Elizabeth S. Allman and John A. Rhodes. Phylogenetic ideals and varieties for the general Markov model, [arXiv:math.AG/0410604](https://arxiv.org/abs/math/0410604), 2004.
- [2] J. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71, 1987.
- [3] James W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.
- [4] J Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, 2004.
- [5] Lior Pachter and Bernd Sturmfels. *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005. To appear.
- [6] Doug Rohde. SVDLIBC. Available at <http://tedlab.mit.edu/~dr/SVDLIBC>.
- [7] David Sankoff and Mathieu Blanchette. Comparative genomics via phylogenetic invariants for Jukes-Cantor semigroups. In *Stochastic models (Ottawa, ON, 1998)*, volume 26 of *CMS Conf. Proc.*, pages 399–418. Amer. Math. Soc., Providence, RI, 2000.
- [8] Chung-I Wu and Wen-Hsiung Li. Evidence for Higher Rates of Nucleotide Substitution in Rodents Than in Man. *PNAS*, 82(6):1741–1745, 1985.