

Computational methods in constructing phylogenetic tree using flat matrix

Young Rock Kim¹, Chun-Jae Park² and Oh-In Kwon and Sung-Hun Paeng¹

1) *Department of Mathematics, Konkuk University, Seoul 143-701, KOREA*

2) *IIRC, Kyunghee University, Suwon, KOREA*

Corresponding Author : Young Rock Kim, rocky777@math.snu.ac.kr

ABSTRACT

Phylogenetic algebraic geometry is concerned with certain complex projective algebraic varieties derived from finite trees. This talk gives a self-contained introduction to this subject and introduces a new, statistically consistent algorithm for phylogenetic tree construction that uses the algebraic theory of statistical models. Our fundamental tool is Singular Value Decomposition (SVD) from numerical linear algebra. Starting with a multiple sequence alignment of n DNA sequences, we show that SVD gives us to decide whether a split of the taxa occurs in their phylogenetic tree, assuming only that evolution follows a tree model. Using this fact, we have developed an algorithm to construct a phylogenetic tree by computing only $O(n^2)$ SVDs relating to modified flat matrix.

INTRODUCTION

Assume that evolution follows a tree model with evolution acting independently at different sites of genome. Let the transition matrices for this model is the general Markov model which is a more general than any other in the Felsenstein hierarchy. Note that statistical models are algebraic varieties, we are interested in the defining polynomials that are called phylogenetic invariants for varieties. Many authors have studied these invariants for different models [Lake, 1987, Cavender and Felsenstein, 1987, Sankoff and Blanchette, 2000, Sturmfels and Sullivant, 2005, Allman and Rodes, 2003, 2004].

The problem with phylogenetic invariants is that there are many polynomials in exponentially many variables to test on exponentially many trees. However, we try to solve combinatorial explosion matter by concentrating on invariants which are given by rank conditions on certain matrices, called flattenings. In this paper we modify this flat matrix a little bit more for computational reason. Next we give the tree-building algorithm, using modified Singular Value Decomposition to calculate how close a matrix is to a certain rank. Also we give several examples for this algorithm. We used the program **seq-gen** to simulate data of various lengths for the tree. After that we align these data via multiple sequence alignment program and use the Neighbor joining algorithm via cherry picking in a tree by defining the distance between species.

We construct the following algorithm.

Tree construction algorithm with modified SVD and flat matrix:

Input: A multiple alignment of genomic data from n species, from the alphabet Σ with $m = 2, 4$ states.

Output: An unrooted binary tree with n leaves labeled by the species.

Initialization: Compute empirical probabilities $p_{i_1 \dots i_n}$. Count occurrences of each possible column of the alignment, ignoring columns with characters not in Σ . Once we compute one of flat matrix, then remember row operations for SVD to reduce computational time.

Loop: For k from n down to 4, operate the following steps.

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

We compare our algorithm with the other algorithm such as PHYLIP, dnaml. Our algorithm constructs trees for more general model comparing to the other methods.

REFERENCES

1. Pavel A. Pevzner,, *Computational Molecular Biology - An algorithmic Approach*, the MIT Press, 2000.
2. Patchter, L. and Sturmfels, B., *Algebraic statistics for computational biology*, Cambridge University Press, 2005.
3. Patchter, L. and Sturmfels, B., “The mathematics of Phylogenomics”, *arXiv:math.ST/0409132v1*, 8 Sep 2004.
4. Eriksson, N., “Tree construction using singular value decomposition”, Chapter 19, *Algebraic statistics for computational biology*, Cambridge University Press, 2005.