Stochastic Models of Molecular Evolution: An Algebraic Analysis

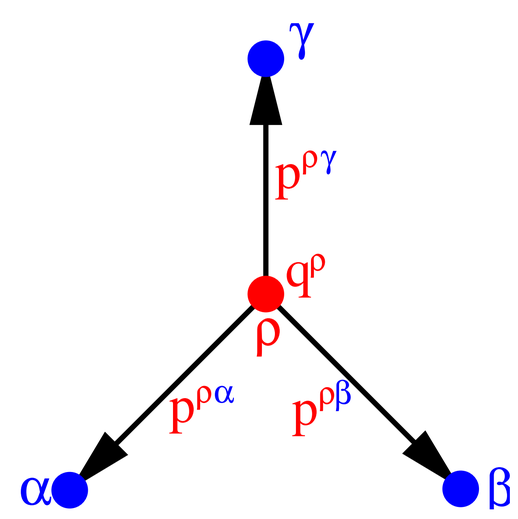
GSF - National Research Centre for Environment and Health
member of the Helmholtz Association

Steffen Kläre
Bioinformatics at Heinrich-Heine-Universität Düsseldorf

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

1 General Model Definitions

Methods of phylogenetic inference derive a (ideally rooted) tree structure $\mathcal{T}_\varrho = (\mathcal{V}, \mathcal{E}, \varrho)$ from a given set of aligned sequences. The Maximum Likelihood method of Felsenstein [1981] employs a Markov process $\mathbf{X} : \mathcal{V} \to \mathcal{S}$. The sequence alignments are treated as information about the process at the leaf set $\mathcal{L}$ of $\mathcal{T}_\varrho$. This information is translated into a joint leaf distribution $\mu^{\mathcal{L}}$. Our goal is to infer conditions on $\mu^{\mathcal{L}}$ under which an underlying Markov process can be recovered. Since triple trees are structurally unique and because Chang [1996] shows that an existing Markov process on a general tree can be reconstructed by looking at its triple restrictions, this work is looking at triple trees only.



The process is characterized by the transition matrices $p^{\varrho\alpha}$, $p^{\varrho\beta}$, $p^{\varrho\gamma}$ and the root distribution q^ϱ. The relationship between the parameters and a vector $\mu = \mu^{\alpha\beta\gamma}$ is given by:

$$\mu^{\alpha\beta\gamma}(x_\alpha, x_\beta, x_\gamma) = \sum_{x_\varrho \in \mathcal{S}} q^\varrho(x_\varrho) p^{\varrho\alpha}(x_\varrho, x_\alpha) p^{\varrho\beta}(x_\varrho, x_\beta) p^{\varrho\gamma}(x_\varrho, x_\gamma). \qquad \text{(L)}$$

This equation system is the basis of further Naturally, system (L) can be extended to larger trees (see e.g. equation (2) in Felsenstein [1981]). Let k denote the number of states.

2 Algebraic Questions and Tools

Inferred Questions

For which vectors $\mu \in \mathbf{C}^{k^3}$ does (L) provide a solution?
How many solutions exist for one vector?
When does such a solution characterize a Markov process?

Used Tools

Consider (L) as the parametrization of the variety of factorizing vectors. Via *polynomial implicitization* this parametrization is reversed to obtain a set of polynomials which characterize the smallest affine variety containing the needed variety. These polynomials are called *phylogenetic invariants* (see e.g. Allman and Rhodes [2003]). Denote the variety identified by the invariants as *phylogenetic variety*. For computations of the invariants the algebra package **Singular** (Greuel et al. [2001]) was used.

Obtained Answers

If a solution exists for a vector μ, then it is in the phylogenetic variety.
The number of needed invariants is bounded from below by the difference of equations and variables $k^3 - (k-1)(2k+1)$.
The vectors with an infinite number of solutions form a zero set in the set of factorizing vectors. The vectors with no solution form a zero set in the phylogenetic variety. The number of solutions for all other vectors is bounded from below by $(k!)$.
These results can be extented to cases with more complex tree structures.

3.1 The General Two State Model

For a vector $\mu^{\alpha\beta\gamma}$ denote by $t_{\delta_1\delta_2}$ the covariance between the leaves δ_1 and δ_2, and by $r_{\delta_1}, \delta \in \mathcal{L}$ and Δ terms in $\mu^{\alpha\beta\gamma}$. If the covariances and Δ are not zero, then the system (L) has a unique solution up to symmetry, which is given by

$$p^{\delta_1}_{00} = -\frac{r_{\delta_1} \pm \sqrt{\Delta}}{2t_{\delta_2\delta_3}}, \quad p^{\delta_1}_{10} = -\frac{r_{\delta_1} \mp \sqrt{\Delta}}{2t_{\delta_2\delta_3}}, \quad q^\varrho_0 = \frac{1}{2} \pm \frac{r_\alpha + 2\mu^\alpha_0 t_{\beta\gamma}}{2\sqrt{\Delta}}.$$

The relationship between the two solutions is visualized by

$$p^{\delta_1}_+ = \begin{pmatrix} \clubsuit & \diamondsuit \\ \heartsuit & \spadesuit \end{pmatrix} \longleftrightarrow \begin{pmatrix} \heartsuit & \spadesuit \\ \clubsuit & \diamondsuit \end{pmatrix} = p^{\delta_1}_-.$$

If the solution characterizes a Markov process then either two leaves are positive correlated with each other and negatively correlated to the third or all three leaves are pairwise positive correlated.
The vectors, for which $\Delta = 0$ but $t_{\alpha\beta}t_{\alpha\gamma}t_{\beta\gamma} \neq 0$ did not return any solution.This observation led to the conjecture that the phylogenetic variety for (L) under the two state model is not affine.

3.2 The Neyman N_k Model

Singular generated two polynomials, the summation condition for probability vectors under the Neyman model and a junk polynomial.

If the Hamming distances $d_{\alpha\beta}$, $d_{\alpha\gamma}$ and $d_{\beta\gamma}$ are not equal to $1 - 1/k$ then the system (L) has a unique solution up to symmetry, which is given by

$$p_{\delta_1} = \frac{1}{k}\left(1 \pm \frac{\Delta}{1 - \frac{k}{k-1}d_{\delta_2\delta_3}}\right),$$

where $\delta = (1 - kd_{\alpha\beta}/(k-1))(1 - kd_{\alpha\gamma}/(k-1))(1 - kd_{\beta\gamma}/(k-1))$. The relationship between these solutions is visualized by

$$\begin{pmatrix} \clubsuit + \spadesuit & \clubsuit & \clubsuit & \clubsuit \\ \clubsuit & \clubsuit + \spadesuit & \clubsuit & \clubsuit \\ \clubsuit & \clubsuit & \clubsuit + \spadesuit & \clubsuit \\ \clubsuit & \clubsuit & \clubsuit & \clubsuit + \spadesuit \end{pmatrix} \longleftrightarrow \begin{pmatrix} \clubsuit - \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} \\ \clubsuit + \frac{\spadesuit}{2} & \clubsuit - \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} \\ \clubsuit + \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} & \clubsuit - \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} \\ \clubsuit + \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} & \clubsuit + \frac{\spadesuit}{2} & \clubsuit - \frac{\spadesuit}{2} \end{pmatrix}$$

Either none, one or both solution describe Markov processes. Note however, that each solutions returns its own vector $\mu^{\mathcal{L}}$.

3.3 The Kimura 2ST Model

Applying **Singular** to the associated form of (L) returned a system of 18 polynomials which filled 24 pages of output.

The three pairwise distributions $\mu^{\alpha\beta}$, $\mu^{\alpha\gamma}$ and $\mu^{\beta\gamma}$ to a given triple leaf distribution $\mu^{\alpha\beta\gamma}$ has four different solutions for (L) if the difference of the probability of staying in one class minus the probability of changing classes and the difference of the probability of no change and a change within a class are both not zero.

$$\begin{pmatrix} \clubsuit & \spadesuit & \diamondsuit & \diamondsuit \\ \spadesuit & \clubsuit & \diamondsuit & \diamondsuit \\ \diamondsuit & \diamondsuit & \clubsuit & \spadesuit \\ \diamondsuit & \diamondsuit & \spadesuit & \clubsuit \end{pmatrix} \longleftrightarrow \begin{pmatrix} \spadesuit & \clubsuit & \diamondsuit & \diamondsuit \\ \clubsuit & \spadesuit & \diamondsuit & \diamondsuit \\ \diamondsuit & \diamondsuit & \spadesuit & \clubsuit \\ \diamondsuit & \diamondsuit & \clubsuit & \spadesuit \end{pmatrix},$$

$$\begin{pmatrix} \clubsuit & \spadesuit & \diamondsuit & \diamondsuit \\ \spadesuit & \clubsuit & \diamondsuit & \diamondsuit \\ \diamondsuit & \diamondsuit & \clubsuit & \spadesuit \\ \diamondsuit & \diamondsuit & \spadesuit & \clubsuit \end{pmatrix} \longleftrightarrow \begin{pmatrix} 2\diamondsuit - \heartsuit & \heartsuit & \frac{\clubsuit + \spadesuit}{2} & \frac{\clubsuit + \spadesuit}{2} \\ \heartsuit & 2\diamondsuit - \heartsuit & \frac{\clubsuit + \spadesuit}{2} & \frac{\clubsuit + \spadesuit}{2} \\ \frac{\clubsuit + \spadesuit}{2} & \frac{\clubsuit + \spadesuit}{2} & 2\diamondsuit - \heartsuit & \heartsuit \\ \frac{\clubsuit + \spadesuit}{2} & \frac{\clubsuit + \spadesuit}{2} & \heartsuit & 2\diamondsuit - \heartsuit \end{pmatrix}.$$

The above matrix changes describe the relationship of the obtained solutions. Note, that the second change also changes the process on the triple tree.

References

Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 186(2):113–144, December 2003.

Joseph T. Chang. Full reconstruction of Markov models on Evolutionary Trees: Identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996.

Joe Felsenstein. Evolutionary Trees from DNA sequences: A Maximum Likelihood approach. *Journal of Molecular Evolution*, 17(6):368–76, 1981.

Gert-Martin Greuel, Gerhard Pfister, and Hans Schönemann. SINGULAR 2.0. A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, University of Kaiserslautern, 2001. `http://www.singular.uni-kl.de`.