

Algebraic Models in Phylogenetics

John A. Rhodes

Dept. of Mathematics and Statistics

University of Alaska Fairbanks

Workshop on Algebraic Statistics
and Computational Biology

Clay Mathematics Institute, November 12-14, 2005

The basic problem of **Molecular Phylogenetics** is to infer evolutionary trees from biological sequence data at the leaves.

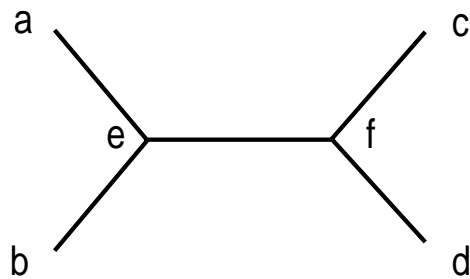
Ex:

a: AATCGTAGCTCGACC...

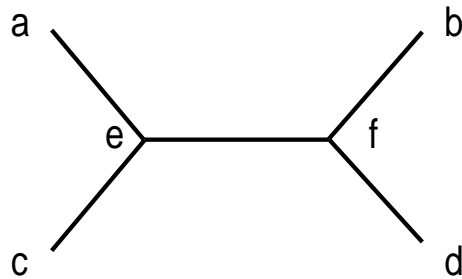
b: AAATGCCGGCTCGACC...

c: AAACGTGACTTGAGC...

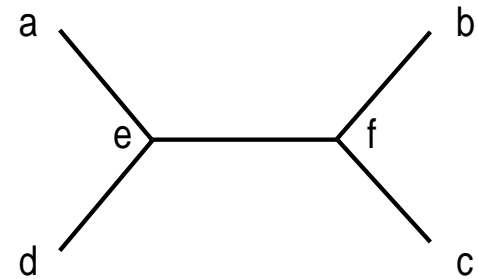
d: AATCGTAGCTTGATC...



T₁



T₂



T₃

Probabilistic models of
sequence evolution.



Polynomial
algebra.

Though the number of polynomials and the number of variables are large, everything is highly structured, due to:

- the underlying tree

and

- biologically-plausible assumptions about the mutation process.

Outline of models of molecular evolution:

ACTGTGTA ACTAACG ...

↓

TCTGTATA ACTCAGG ...

Model **one site** — each site in data sequences is a trial of the same process, **i.i.d.**

There are κ possible states for a site

- $\kappa = 4$, states A, G, C, T ,
 - $\kappa = 2$, states $R = \{A, G\}$, $Y = \{C, T\}$, (purines, pyrimidines)
- or
- $\kappa = 20$ states, the amino acids of protein sequences



Root distribution vector:

$$\mathbf{p}_r = (p_i) = (p_A \ p_G \ p_C \ p_T) = (.25 \ .25 \ .25 \ .25)$$

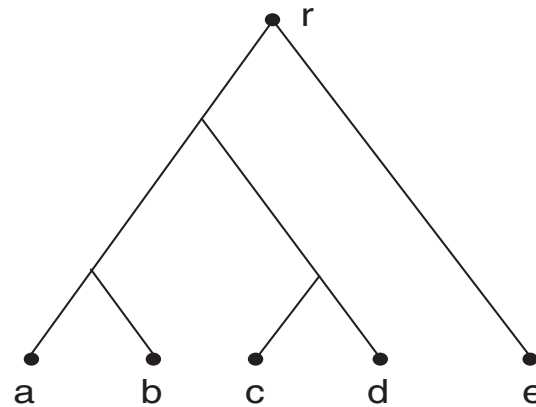
Markov matrix of conditional probabilities of substitutions:

$$M = (p(j|i)) = (\text{Prob}(j \text{ at } a \mid i \text{ at } r))$$

$$= \begin{pmatrix} p(A|A) & p(G|A) & p(C|A) & p(T|A) \\ p(A|G) & p(G|G) & p(C|G) & p(T|G) \\ p(A|C) & p(G|C) & p(C|C) & p(T|C) \\ p(A|T) & p(G|T) & p(C|T) & p(T|T) \end{pmatrix} = \begin{pmatrix} .6 & .2 & .1 & .1 \\ .2 & .6 & .1 & .1 \\ .1 & .1 & .6 & .2 \\ .1 & .1 & .2 & .6 \end{pmatrix}$$

But evolution occurs along a tree.

Ex: 5 taxa

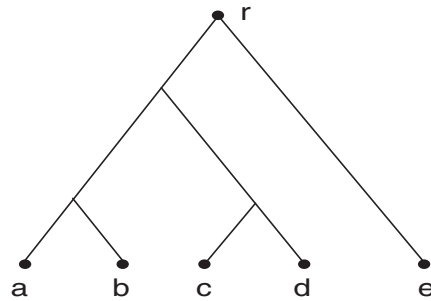


Model parameters:

- tree T (105 possibilities),
- root base distribution vector, $\mathbf{p} = (p_A p_G p_C p_T)$,
- Markov matrices, M_e , for each edge,

so $3 + 8(12) = 99$ dimensional parameter space for $\kappa = 4$.

We observe sequences only of living taxa — at leaves of tree;



Joint dist. P at leaves is a $4 \times 4 \times 4 \times 4 \times 4$ tensor (1024 entries),

E.g., $P(1, 1, 3, 4, 4) = P(A, A, C, T, T) =$ probability of

a: ...A...

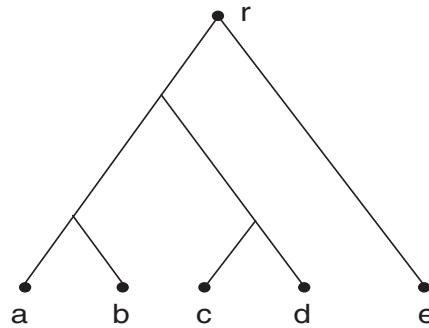
b: ...A...

c: ...C...

d: ...T...

e: ...T...

The joint distribution P is a function of the parameters:



$$\begin{aligned}
 P(i, j, k, l, m) = & \sum_{s=1}^4 \sum_{t=1}^4 \sum_{u=1}^4 \sum_{v=1}^4 p_s M_1(s, m) \cdot \\
 & M_2(s, t) M_3(t, u) M_4(u, l) M_5(u, k) \cdot \\
 & M_6(t, v) M_7(v, j) M_8(v, i),
 \end{aligned}$$

9th degree, 99 variables, 256 terms, reflecting the topology of T.

Basic Inference Problem:

From sequence data, estimate P by the observed distribution \hat{P} .

From \hat{P} , infer the unknown parameters of the model

— especially the tree T .

For fixed choice of T , parameterization map of all possible joint distributions P :

$$\phi_T : [0, 1]^{99} \rightarrow [0, 1]^{1024},$$

a polynomial map, which by the same formulas extends to

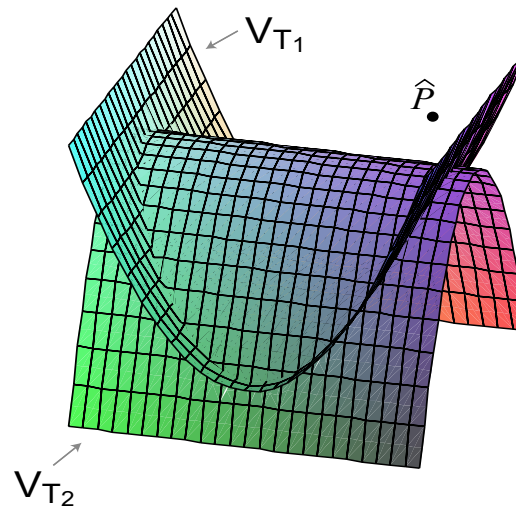
$$\phi_T : \mathbb{C}^{99} \rightarrow \mathbb{C}^{1024}.$$

The **phylogenetic variety for T** is (the closure) of the image of this map:

$$V_T = \overline{\phi_T(\mathbb{C}^{99})}.$$

the variety of all joint distributions arising from the model on T
(allowing complex parameters)

Informally, V_T is the collection of all joint distributions that could arise from T , considering all possible numerical parameters



Tree Inference:

Given the observed distribution \hat{P} , an estimate of the expected distribution P arising from an unknown T , determine T so that \hat{P} is 'closest' to V_T .

The model described so far is the **general Markov model**, but there are many interesting biologically/mathematically-motivated variants:

- restriction of Markov matrices to special forms

Group-based models, e.g.,
$$\begin{pmatrix} * & a & b & c \\ a & * & c & b \\ b & c & * & a \\ c & b & a & * \end{pmatrix}$$

Strand-symmetric model

Stable base-distribution models

Time-reversible models

- variation of mutation rates between sites/across trees

Mixture models

Covarion model

-

$V_T = \overline{\text{Im}(\phi_T)}$ can also be described as the zero-set of a collection of polynomials, I_T , the **phylogenetic ideal**.

$$f \in I_T \Leftrightarrow f(P) = 0 \text{ for all } P \in \text{Im } \phi(T)$$

Polynomials in I_T are **phylogenetic invariants**, which

- are constraints on the form of the distributions arising from T and the model,
- describe V_T implicitly.

Ex: For $\kappa = 4$, 5-taxon tree, $f \in I_T$ is a polynomial in 4^5 entries of P , reflecting the structure of T and the model.

Phylogenetic invariants were proposed for use in tree inference in 1987 (Cavender-Felsenstein, Lake).

Two main questions:

- For a fixed model, how do we find (some, most, all) phylogenetic invariants for all trees T ?

Naive answer: It's just a computation.

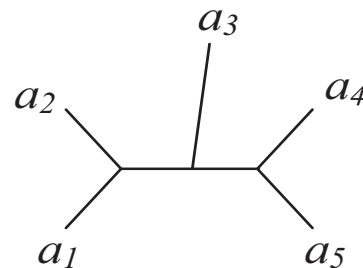
- Once we have found invariants, how do we use them?

Naive answer: If $f \in I_T$, then $f(\hat{P}) \approx 0$ is evidence in support of T being the tree behind the data.

Both questions offer opportunities for research.

I: Finding invariants

Ex: $\kappa = 2$ states; P a $2 \times 2 \times 2 \times 2 \times 2$ tensor,



P has two natural *flattenings* according to *splits* in the tree:

$\{\{a_1, a_2\}, \{a_3, a_4, a_5\}\}$, and $\{\{a_1, a_2, a_3\}, \{a_4, a_5\}\}$.

The corresponding flattenings are

$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}$$

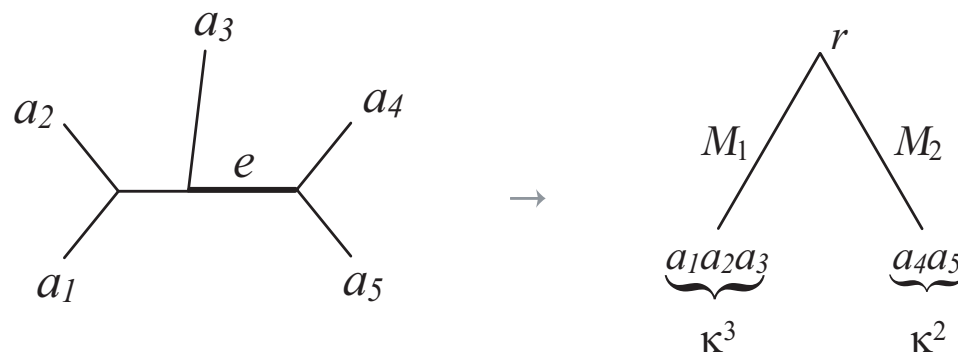
and

$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}.$$

Claim: All 3×3 minors of these two matrices are phylogenetic invariants for GM model on T , and they generate I_T .

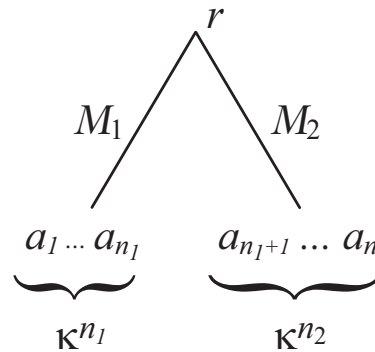
Why are these invariants?

Associated to a flattening of P is a *coarsening* of the model:



$$\begin{array}{ccc}
 P & \mapsto & \text{Flat}_e(P) \\
 2 \times 2 \times 2 \times 2 \times 2 & & 2^3 \times 2^2
 \end{array}$$

Coarsened model:



where M_1 is 2×2^{n_1} , M_2 is 2×2^{n_2} , with entries polynomial in original parameters, and

$$\text{Flat}_e(P) = M_1^T \text{diag}(\mathbf{p}_r) M_2.$$

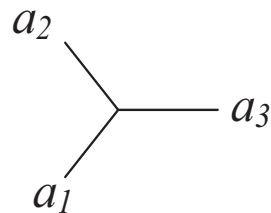
Thus $\text{Flat}_e(P)$ has rank 2, so 3×3 minors vanish.

Theorem (Allman, R 2005): If $\kappa = 2$ and T bifurcating, then I_T is generated by the *edge invariants*, the 3×3 minors of edge flattenings.

The proof that these generate the ideal involves other ideas: group actions on varieties, representations of GL_4 , special features of $\kappa = 2$.

$\kappa > 2$, current results are less complete...

Ex: 3 taxa, $\kappa = 4$ states (DNA)



There are no edge invariants for this tree, but

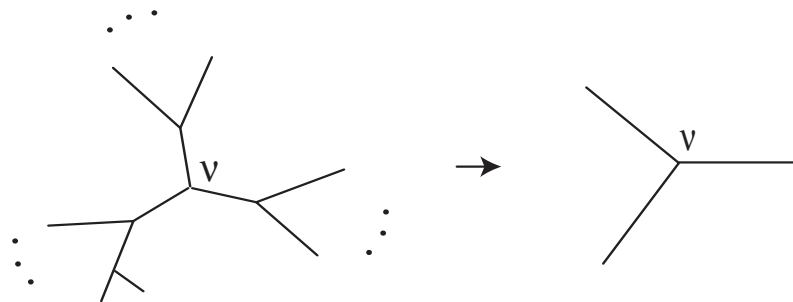
A 1728-dim space of all quintic polynomials in I_T can be explicitly constructed, ... but ideal generators are not known.

Such **vertex invariants** are related to **3-dim tensor rank**,
much as **edge invariants** are related to **matrix rank**.

(For algebraic geometers, $V_T = \text{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$.)

Vertex invariants for an arbitrary tree:

Focus on an internal node, coarsen model



For $P \in V_T$, flatten

$$P \quad \mapsto \quad \text{Flat}_v(P)$$
$$4 \times 4 \times \dots \times 4 \quad 4^m \times 4^l \times 4^k$$

Theorem (Allman, R 2005): For any κ , if an explicit set of polynomials defining V_T for the 3-leaf tree is known, then for any bifurcating T explicit polynomials defining V_T can be given.

Notes:

- 1) This only concerns finding polynomials whose zero set is V_T ; it does not claim these generate the full ideal.
- 2) The polynomials are naturally related to edges and vertices, the local features of the tree.

For other models:

Group-based models: Full ideal is known

through Fourier transform on (abelian) group (Hendy 1986, Evans-Speed 1993, Székeley-Steel-Erdős 1993)

and recognition that variety is toric (Sturmfels-Sullivant 2005).

Strand-Symmetric model: 'most' invariants are known

4-state model amalgamating a 2-state group-based and 2-state general Markov (Cassanellas-Sullivant 2005)

Other model results are much less complete.

II: Using invariants

A little history...

- Much early work focused on **linear** invariants for various models.
- Linearity implies invariants still vanish when applied to **mixture models** with **across-site rate variation**.
- Vanishing of linear invariants alone can give a **statistically consistent** method of tree inference ...
- ... but simulation studies showed **poor performance** on short sequences — worse than other approaches.

- ... but poor performance was only using **only linear** invariants

With all invariants, there is reason to be optimistic:

1) Simulation studies with group-based models on small trees (Cassanellas-Garcia-Sullivant 2005) indicate good performance on shorter sequences

2) Since

phylogenetic distance

+

4-point condition

= specific invariant,

clever use of invariants should at least match distance methods

More recently,

- New Tree construction algorithm

Eriksson (2005), based on edge invariants for GM

Edge invariants \leftrightarrow matrix rank conditions \leftrightarrow approx. rank via SVD

- 1) Compute observed distribution \hat{P} for sequences from n taxa, and create a node for each taxon.
- 2) For each possible split of nodes into partition of size 2, $(n-2)$, compute approximate rank of corresponding flattening of \hat{P} .
- 3) Join the 'best' choice of 2 nodes to a new common node, reducing the possible choices at the next stage to $n - 1$ nodes.

Performance of this algorithm is good (not great) in simulation tests, though tests probably biased against it.

There is much potential to improve it, with work continuing.

- Potential use of invariants to measure support for edges/nodes

For some biological issues existence of a clade may be more interesting than its precise resolution.

With full ML, no good means of isolating support for a single feature in a tree (e.g, bootstrapping)

Can a meaningful statistical measure of support be developed from invariants?

Current project: Are invariants useful for heuristic searching of tree space to find ML tree?

Tree-space search uses tree-bisection-and-reconnection (TBR) steps; choose bisection points using invariants.

- Understanding exact solution of ML

Knowledge of invariants allows the ML problem to be phrased as a constrained optimization problem – invariants provide the constraints.

For small trees, simple models, Chor-Hendy-Holland-Penny (2000) showed there can be

- multiple local optima
- multiple global optima
- even a continuum of global optima

A molecular clock hypothesis can lead to unique optimum
Chor-Khetan-Snir (2003)

- Identifiability of Trees

Basic question: Given a joint distribution P known to have arisen from a particular model, is it possible to recover all the parameters? If so, the model is **identifiable**.

For a specific model, is the tree parameter T identifiable?

- If the answer is no, we cannot infer trees reliably even with ‘perfect’ data.
- For sufficiently complicated models, trees are not identifiable.
(e.g., for mixture models with m classes provided $\text{Sec}^m(V_T)$ fills space.)

For many models (GM, group-based, GTR+ Γ +I) tree identifiability has been proven (usually using distances and the 4-point condition)

For general **mixture** or **covarion** models, the question was open.

Mixture models:

Ex:

GM+GM : 2 classes of sites, each mutating according to different GM parameters on the same tree. Which site is in which class is unknown, size of classes is unknown.

GM+GM+I : 3 classes of sites, one invariable (modeling functional constraints on gene).

Covarion model (Tuffley-Steel 1998)

Sites that are invariable in one part of the tree may become variable in another, and *vice versa*.

To model this, need 8 states at internal nodes:

$$A^{\text{on}}, C^{\text{on}}, G^{\text{on}}, T^{\text{on}}, A^{\text{off}}, C^{\text{off}}, G^{\text{off}}, T^{\text{off}}$$

but only 4 observable states at leaves:

$$A, C, G, T$$

Allowing sites to switch between variable/invariable modes in different parts of tree is believed to increase biological realism, especially for highly divergent taxa.

When sites 'on,' mutation described by

- a 4-element base distribution vector \mathbf{p} ,
- a 4×4 rate matrix R (time-reversible), with $\mathbf{p}R = \mathbf{0}$.

Also have 'on/off' switching rate parameters s_1, s_2 .

Then

$$Q = \begin{pmatrix} R - s_1 I & s_1 I \\ s_2 I & -s_2 I \end{pmatrix}$$

is the rate matrix for a 8-state time-reversible process, stationary on the root distribution vector $\tilde{\mathbf{p}} = \left(\frac{s_2}{s_1 + s_2} \mathbf{p}, \frac{s_1}{s_1 + s_2} \mathbf{p} \right)$.

$$M_e = \exp(Q t_e), \text{ where } t_e \text{ is a scalar edge length}$$

Theorem (Allman, R 2005): The tree topology is identifiable for generic parameters of the covarion model if $\kappa \geq 3$.

The proof considers a more general model on a tree T , with purely algebraic definition (no common rate matrix).

(λ, κ) -state general Markov model: $\lambda \geq \kappa$

λ states at all internal nodes,

κ states at leaves.

Parameters:

- root distribution vector \mathbf{p} with λ entries
- $\lambda \times \lambda$ Markov matrices on all internal edges
- $\lambda \times \kappa$ Markov matrices on terminal edges

For $\kappa < \lambda < \kappa^2$, **some** phylogenetic invariants are constructed for this model.

Vanishing of these invariants on $P \in \text{Im}(\phi_T)$ is enough to determine T (for generic numerical parameters)

With identifiability of (λ, κ) -model proved for $\kappa \leq \lambda < \kappa^2$, must still specialize to covarion model (with $\lambda = 8, \kappa = 4$)

— and show specialization is generic.

Also, (λ, κ) -model specializes in other ways, yielding new identifiability results for many other models:

- GM: $\lambda = \kappa$,
- GM+I: $\lambda = 2\kappa$, $\begin{pmatrix} M & 0 \\ 0 & I \end{pmatrix}$,
- GM+GM+...+GM model ($m < \kappa$ summands): $\lambda = m\kappa$
- rates-across-sites models with $< \kappa$ arbitrary rate classes

Note: These invariants give both **theoretical** and potentially **practical** means of identifying tree topologies.

References:

- E. S. Allman and J. A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.
- Phylogenetic invariants for stationary base composition. *J. Symbolic Comp.*, 2005, to appear, arXiv:q-bio.PE/0407035.
- Phylogenetic ideals and varieties for the general Markov model. 2004, preprint, arXiv:math.AG/0410604.
- The identifiability of tree topology for phylogenetic models, including covarion and mixture models. 2005, preprint, arXiv:q-bio.PE/0511009.
- Phylogenetic invariants and parameter recovery for the general Markov plus invariable sites model. 2005, in preparation.
- M. Casanellas and L.D. Garcia and S. Sullivant. Catalog of Small Trees. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005.
- M. Casanellas and S. Sullivant. The strand symmetric model. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005.
- J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. of Class.*, 4:57–71, 1987.
- B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Bio. and Evol.*, 17:1529–1541, 2000.

B. Chor, A. Khetan, and S. Snir. Maximum likelihood on Four Taxa Phylogenetic Trees: analytic solutions. RECOMB'03. 2003.

N. Eriksson. Tree construction using singular value decomposition. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*. CUP, 2005.

S. N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

M. D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, 38:310–321, 1989.

M. D. Hendy and D. Penny. A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38:297–309, 1989.

J. A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Bio. Evol.*, 4:167–191, 1987.

L. Pachter and B. Sturmfels, editors. *Algebraic statistics for computational biology*. Cambridge University Press, 2005.

L. A. Székely, M. A. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. *Adv. in Appl. Math.*, 14(2):200–210, 1993.

B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 2005. arXiv:q-bio.PE/0402015.

Chris Tuffley and Mike Steel. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.*, 147(1):63–91, 1998.