FOURIER ANALYSIS AND PHYLOGENETIC TREES

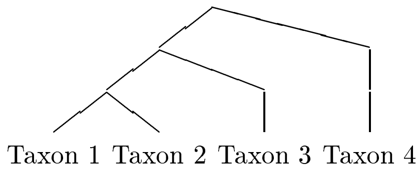
STEVEN N. EVANS

Abstract. We give an overview of phylogenetic invariants: a technique for reconstructing evolutionary family trees from DNA sequence data. This method is useful in practice and is based on a number of simple ideas from elementary group theory, probability, linear algebra, and commutative algebra.

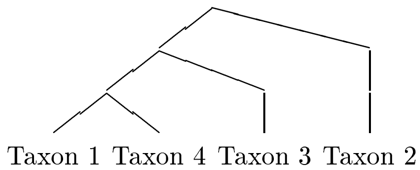
1. Introduction

Phylogeny is the branch of biology that seeks to reconstruct evolutionary "family trees." Such reconstruction can take place at various scales. For example, we could attempt to build the family tree for various present day indigenous populations in the Americas and Asia in order to glean information about the possible course of migration of humans into the Americas. At the level of species, we could seek to determine whether modern humans are more closely related to chimpanzees or to gorillas. Ultimately, we would like to be able to reconstruct the entire "tree of life" that describes the course of evolution leading to all present day species. Because the status of the "leaves" on which we wish to build a tree differs from instance to instance, biologists use the general term *taxa* (singular *taxon*) for the leaves in a general phylogenetic problem.

For example, for 4 taxa, we might seek to decide whether this tree



or this tree



describes the course of evolution. In such trees:

- the arrow of time is down the page,
- paths down through the tree represent *lineages* (*lines of descent*),

Date: March 4, 2002.
1991 *Mathematics Subject Classification.* Primary: 62P10, 13P10. Secondary: 68Q40, 20K01.
Key words and phrases. invariant, phylogeny, DNA, genome, tree, discrete Fourier analysis, algebraic variety, elimination ideal, free module.
Research supported in part by NSF grant DMS-0071468.

- any point on a lineage corresponds to point of time in the life of some ancestor of a taxon,
- vertices other than leaves represent times at which lineages diverge (that is, times at which certain taxa cease to have common ancestors),
- the root corresponds to the most recent common ancestor of all the taxa.

Phylogenetic reconstruction has a long history. Classically, reconstruction was based on the observation and measurement of morphological similarities between taxa with the the possible adjunction of similar evidence from the fossil record; and these methods continue to be used. However, with the recent explosion in technology for sequencing large pieces of a genome rapidly and cheaply, reconstruction from the huge amounts of readily available DNA sequence data is now by far the most commonly used technique. Moreover, reconstruction from DNA sequence data has the added attraction that it can operate fairly automatically on quite well-defined digital data sets that fit into the framework of classical statistics, rather than proceeding from a somewhat ill-defined mix of qualitative and quantitative data with the need for expert oversight to adjust for difficulties such as morphological similarity due to convergent evolution.

There is a substantial literature on both the mathematics behind various approaches to phylogenetic reconstruction and the algorithmic issues that arise when we try to implement these approaches with large amounts of data and large numbers of taxa. We won't attempt to survey this literature or provide a complete bibliography. Rather, these lecture notes are devoted to some of the mathematics behind one particular approach: that of *phylogenetic invariants*. Not only is this technique of practical utility, but it requires a nice combination of elementary group theory, probability, linear algebra, and commutative algebra.

The outline of the rest of these notes is as follows. Section 2 begins with a discussion of the sort of DNA sequence data that are used for phylogenetic reconstruction and how these data are pre-processed using sequence alignment techniques. We then describe a very general class of "Markov random field" models that incorporate arbitrary mechanisms for nucleotide substitution and a dependence structure for the nucleotides exhibited by the taxa that mirrors the phylogenetic tree. Section 3 introduces 3 restricted classes of substitution mechanisms that are commonly used in the literature: the Jukes-Cantor model and the 2- and 3-parameter Kimura models. We observe in Section 4 that standard statistical techniques such as maximum likelihood are still computationally very demanding for infering phylogenies even for such restricted models and we propose the alternative approach of phylogenetic invariants. We point out in Sections 5 and 6 that an underlying group structure is present in the restricted substitution models and develop the Fourier analysis that is necessary for exploiting this group structure to construct and recognise invariants.

Section 7 is a warm-up that uses these algebraic tools to exhibit an invariant for a particular tree. The ideas in this section are then generalised in Section 8 to characterise the class of all invariants for an arbitrary tree. Finally, we determine the "dimension" of the space of invariants for an arbitrary tree in Section 9 and show in Section 10 that different trees have different invariants, with the "dimension" of the class of distinguishing invariants depending in a simple manner on the difference between the two trees.

2. Data and general models

We assume that reader is familiar with the basic notion of the hereditary information of organisms being carried by DNA molecules that consist of two linked chains built from an alphabet of four *nucleotides* and twisted around each other in a double helix, and, moreover, that such a molecule can be described by listing the sequence of the nucleotides encountered along one of the chains using the letters A=adenine, G=guanine, C=cytosine, T=thymine. A lively and entertaining guide to the fundamentals is [GW91].

The totality of the DNA in any somatic cell constitutes the genome of the individual. The genomes of different individuals differ. As evolution occurs, one nucleotide is substituted for another, segments of DNA are deleted, and new segments are inserted.

Sequence alignment is a procedure that attempts to provide algorithms that takes DNA sequences from several taxa, line up "common positions" at which substitutions may or may not have occurred, and determine where deletions and insertions have occurred in certain sequences relative to the others. For example, an alignment of two taxa might produce an output such as the following:

$$\begin{array}{ll} \text{Taxon 1} & ...AGTAACT... \\ \text{Taxon 2} & ...AT***CA... \end{array}$$

Reading from left to right: both taxa have an A in the "same" position, the next position is common to both taxa but Taxon 1 has a G there whereas Taxon 2 has a T, then (due to insertions or deletions) there is a stretch of 3 positions that are present in the genome of Taxon 1 but not present in the genome of Taxon 2 *etc.* There are many approaches to deriving such alignments, and a discussion of them is outside the scope of these notes. A good introduction to some of the mathematical issues is [Wat95].

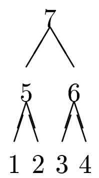
Our basic data are DNA sequences for each of our taxa that have been preprocessed in some suitable way to align them. For simplicity, we suppose that we are dealing with segments where there have been no insertions or deletions, so all the taxa share the same common positions and differences between nucleotides at these positions are due to substitutions.

The standard statistical paradigm dictates (in very broad terms) how we should go about taking these data and producing inferences about the phylogeny connecting our taxa. Firstly, we should begin with a probability model that incorporates the possible trees as a "parameter" along with other parameters that describe the mechanism by which substitutions occur relative to such a tree. Secondly, we should determine the choice of parameters (in particular, the choice of tree) that best fits the observed sequence data according to some criterion.

A standard assumption in the literature is that the behaviour at widely separated positions on the genome is statistically independent. With this assumption, the modelling problem reduces to one of modelling the nucleotide observed at a given position.

In order to describe the general class of single position models typically used in the literature, it is easiest to begin by imagining that we can observe not only the nucleotides for the taxa but also those for the unobserved intermediates represented by the interior vertices of the tree. (For simplicity, let us refer to the taxa and the intermediates as "individuals" for the moment.) Two individuals share the same lineage up to their most recent common ancestor and so the processes such as mutation leading to substitution act on the genomes of their common ancestors in the same way up until the split in lineages that occurs at the most recent common ancestor. After the split in lineages, it is a reasonable first approximation to assume that the random mechanisms by which substitutions occur are operating independently on the genomes of the ancestors that are no longer shared. Mathematically, this translates into an assumption that that the nucleotides exhibited by two individuals are conditionally independent given the nucleotide exhibited by their most recent common ancestor. Equivalently, the nucleotides exhibited by two individuals are conditionally independent given the nucleotide exhibited by **any** individual on the path that connects the two individuals in the tree.

For example, consider the 4 taxa tree



Letting Y_i denote the nucleotide exhibited by individual i, we have, for example, that

- Y_1 and Y_2 are conditionally independent given Y_5,
- the pair (Y_1, Y_2) are conditionally independent of the pair (Y_3, Y_4) given any one of Y_5, Y_6, or Y_7.

Because of this dependence structure, a joint probability such as

$$\mathbb{P}\{Y_1 = A, Y_2 = A, Y_3 = G, Y_4 = C, Y_5 = T, Y_6 = T, Y_7 = A\}$$

can be computed as

$$\begin{aligned}
&\mathbb{P}\{Y_7 = A\} \\
&\quad \times \mathbb{P}\{Y_5 = T \mid Y_7 = A\}\mathbb{P}\{Y_6 = T \mid Y_7 = A\} \\
&\qquad \times \mathbb{P}\{Y_1 = A \mid Y_5 = T\}\mathbb{P}\{Y_2 = A \mid Y_5 = T\} \\
&\qquad \times \mathbb{P}\{Y_3 = G \mid Y_6 = T\}\mathbb{P}\{Y_4 = C \mid Y_6 = T\}.
\end{aligned}$$

Thus, for a given tree, the joint probabilities of the individuals exhibiting a particular set of nucleotides are determined by the vector of 4 unconditional probabilities for the root individual and the 4×4 matrices of conditional probabilities for each edge.

Given such a model for the nucleotides exhibited by all the individuals (taxa and intermediates), we obtain a model for the nucleotides exhibited by the taxa by taking the marginal probability distribution for the taxa. Operationally, this just means that we sum over the possibilities for the intermediates.

For example, suppose that we have the 2 taxa tree

Then, for example,

$$\begin{aligned}\mathbb{P}\{Y_1 = A, Y_2 = G\} &= \mathbb{P}\{Y_1 = A, Y_2 = G, Y_3 = A\} + \mathbb{P}\{Y_1 = A, Y_2 = G, Y_3 = G\} \\ &\quad + \mathbb{P}\{Y_1 = A, Y_2 = G, Y_3 = C\} + \mathbb{P}\{Y_1 = A, Y_2 = G, Y_3 = T\} \\ &= \mathbb{P}\{Y_3 = A\}\mathbb{P}\{Y_1 = A \,|\, Y_3 = A\}\mathbb{P}\{Y_2 = G \,|\, Y_3 = A\} \\ &\quad + \mathbb{P}\{Y_3 = G\}\mathbb{P}\{Y_1 = A \,|\, Y_3 = G\}\mathbb{P}\{Y_2 = G \,|\, Y_3 = G\} \\ &\quad + \mathbb{P}\{Y_3 = C\}\mathbb{P}\{Y_1 = A \,|\, Y_3 = C\}\mathbb{P}\{Y_2 = G \,|\, Y_3 = C\} \\ &\quad + \mathbb{P}\{Y_3 = T\}\mathbb{P}\{Y_1 = A \,|\, Y_3 = T\}\mathbb{P}\{Y_2 = G \,|\, Y_3 = T\}.\end{aligned}$$

We now introduce some notation to describe in full generality the sort of model we have just outlined.

Let $\mathbf{T}$ be a finite rooted tree. Write ρ for the root of $\mathbf{T}$, $\mathbf{V}$ for the set of vertices of $\mathbf{T}$, and $\mathbf{L} \subset \mathbf{V}$ for the set of leaves. We regard $\mathbf{T}$ as a directed graph with edge directions leading away from the root. The elements of $\mathbf{L}$ correspond to the taxa, the tree $\mathbf{T}$ is the phylogenetic tree for the taxa, and the elements of $\mathbf{V}\backslash\mathbf{L}$ correspond to ancestors alive at times when the lineages of taxa diverge. It is convenient to enumerate $\mathbf{L}$ as $(l_1, \dots, l_m)$ and $\mathbf{V}$ as $(v_1, \dots, v_n)$, with the convention that $l_j = v_j$ for $j = 1, \dots, m$ and $\rho = v_n$.

Each vertex $v \in \mathbf{V}$ other than the root ρ has a a *father* $\sigma(v)$ (that is, there is a unique $\sigma(v) \in \mathbf{V}$ such that the directed edge $(\sigma(v), v)$ is in the rooted tree $\mathbf{T}$.) If v_α and v_ω are two vertices such that there exist vertices $v_\beta, v_\gamma \dots, v_\xi$ with $\sigma(v_\beta) = v_\alpha$, $\sigma(v_\gamma) = v_\beta$, $\dots, \sigma(v_\omega) = v_\xi$ (that is, there is a directed path in $\mathbf{T}$ from α to ω), then we say that v_ω is a descendent of v_α or that v_α is an ancestor of v_ω and we write $v_\alpha \le v_\omega$ or $v_\omega \ge v_\alpha$. Note that a vertex is its own ancestor and its own descendent. The *outdegree* $\mathrm{outdeg}(u)$ of $u \in \mathbf{V}$ is the number of *children* of u, that is, the number of $v \in \mathbf{V}$ such that $u = \sigma(v)$. To avoid degeneracies we always suppose that $\mathrm{outdeg}(v) \ge 2$ for all $v \in \mathbf{V}\backslash\mathbf{L}$. (**Note:** Terms such as "father" and "child" are just standard terminology from the theory of trees and don't have any biological significance — an edge in our tree may correspond to thousands of actual generations.)

Let π be a probability distribution on $\{A, G, C, T\}$ – the *root distribution*, The probability $\pi(B)$ is the probability that the common ancestor at the root exhibits nucleotide B. For each vertex $v \in \mathbf{V}\backslash\{\rho\}$, let $P^{(v)}$ be a stochastic matrix on $\{A, G, C, T\}$ (that is, the rows of $P^{(v)}$ are probability distributions on $\{A, G, C, T\}$.) We refer to $P^{(v)}$ as the *substitution matrix* associated with the edge $(\sigma(v), v)$. The entry $P^{(v)}(B', B'')$ is the conditional probability that the individual at vertex v exhibits nucleotide B'' given that the individual at vertex $\sigma(v)$ exhibits nucleotide $B' \in \{A, G, C, T\}$.

Define a probability distribution μ on $\{A, G, C, T\}^{\mathbf{V}}$ by setting

$$\mu((B_v)_{v \in \mathbf{V}}) := \pi(B_\rho) \prod_{v \in \mathbf{V}\backslash\{\rho\}} P^{(v)}(B_{\sigma(v)}, B_v).$$

The distribution μ is the joint distribution of the nucleotides exhibited by all of the individuals in the tree, both the taxa and the unobserved ancestors. The induced marginal distribution on $\{A, G, C, T\}^{\mathbf{L}}$ is

$$p((B_\ell)_{\ell \in \mathbf{L}}) := \sum_{v \in \mathbf{V}\backslash\mathbf{L}} \sum_{B_v} \mu(((B_v)_{v \in \mathbf{V}\backslash\mathbf{L}}, (B_\ell)_{\ell \in \mathbf{L}})),$$

where each of the dummy variables B_v, $v \in \mathbf{V}\backslash\mathbf{L}$, is summed over the set $\{A, G, C, T\}$. The distribution p is the joint distribution of the nucleotides exhibited by the taxa.

With this model in hand, we could try to make inferences from sequence data using standard statistical techniques. For example, we could apply the method of maximum likelihood where we determine the choice of the parameters $\mathbf{T}$, π, and $P^{(v)}$, $v \in \mathbf{V}\backslash\{\rho\}$, that makes the probability of the observed data greatest. (As we discussed above, we would need to observe the nucleotides at several positions and assume they were independent and governed by the same single–position model.) Maximum likelihood is known to have various optimality properties when we have large numbers of data, but unless we have just a few taxa there are a huge number of parameters over which we have to optimise and implementing maximum likelihood directly is numerically infeasible. There are various approaches to overcoming these difficulties – for instance, we can maximise likelihoods 4 taxa at a time and hope to fit the subtrees inferred in this manner into one overall tree for all the taxa. Another approach is to constrain the substitution matrices in some way and hope that the extra structure this introduces makes the inferential problem easier to solve (while still retaining some degree of biological plausibility.) That is the approach we will follow starting in the next section.

3. More specific models

The general model for the observed nucleotides outlined in the Section 2 allows the substitution matrices to be arbitrary. As we discussed in the Section 2, there are practical reasons for constraining the form of these matrices.

The substitution matrix $P^{(v)}$ represents the cumulative effect of the substitutions that occur between the times that the individuals associated with $\sigma(v)$ and v were alive. In order to arrive at a reasonable form for $P^{(v)}$, it is profitable to think about how we would go about modelling the dynamics of this substitution process.

The most natural and tractable dynamics are (time-homogeneous) Markovian ones. That is, if the position currently exhibits a certain nucleotide, B' say, then (independently of the past) the nucleotide changes at rate $r(B', B'')$ to some other nucleotide B''. More formally, if the position currently exhibits nucleotide B', then:

- independently of the past, the probability that the elapsed time until a change occurs is greater than t is $\exp(-\sum_{B''} r(B', B'')\, t)$,
- independently of how long it takes until a change occurs, the probability that it is to B'' is proportional to $r(B', B'')$.

There are obvious caveats in the use of such Markov chain models. Certain positions on the genome can't be altered without serious consequences for the viability of the organism, and so a model that allows substitution to occur in a completely random fashion is not appropriate at such positions. However, if we look at positions that are not associated with regions of the genome that have an identifiable function, then it is somewhat difficult to recognise two positions as being the "same" in two different individuals for the purposes of alignment. Some care is therefore necessary in practice to find positions that can be aligned but are such that a Markov chain model is implausible.

The simplest Markov chain model for nucleotide substitution is the Jukes-Cantor model [JC69, Ney71] in which $r(B', B'')$ is the same for all B', B''. Under this model, the distribution of the amount of time spent at a nucleotide before a change occurs does not depend on the nucleotide and all 3 choices of the new nucleotide are equally likely when a change occurs.

Biochemically, the nucleotides fall into two families: the *purines* (adenine and guanine) and the *pyrimidines* (cytosine and thymine). Substitutions within a family are called *transitions*, and they have a different biochemical status to substitutions between families, which are called *transversions*. Kimura [Kim80] proposed a model that recognised this distinction by assigning a common rate to all the transversions and possibly different common rate to all the transitions. We can represent the rates schematically as follows:

$$\begin{tikzcd} A \arrow[r, dashed, leftrightarrow] \arrow[d, leftrightarrow] \arrow[dr, dashed, leftrightarrow] & C \arrow[d, leftrightarrow] \arrow[dl, dashed, leftrightarrow] \\ G \arrow[r, dashed, leftrightarrow] & T \end{tikzcd}$$

The solid arrows represent transitions and the dashed arrows represent transversions. There are two rate parameters, $\alpha, \beta > 0$, say, such that $r(B', B'') = \alpha$ (resp. $r(B', B'') = \beta$) if B' and B'' are connected by a solid (resp. dashed) arrow.

Later, Kimura [Kim81] introduced a generalisation of this model with the following rate structure:

$$\begin{tikzcd} A \arrow[r, dashed, leftrightarrow] \arrow[d, leftrightarrow] \arrow[dr, Leftrightarrow] & C \arrow[d, leftrightarrow] \arrow[dl, Leftrightarrow] \\ G \arrow[r, dashed, leftrightarrow] & T \end{tikzcd}$$

Now there are 3 types of arrows (solid, dashed, and double) and 3 corresponding rate parameters ($\alpha, \beta, \gamma > 0$, say.) For example, if the current nucleotide is A then, independently of the past, the probability that it takes longer than time t until a change is $\exp(-(\alpha + \beta + \gamma)t)$ and, independently of how long it takes until a change, the change is to G with probability $\alpha/(\alpha + \beta + \gamma)$, to C with probability $\beta/(\alpha + \beta + \gamma)$, and to T with probability $\gamma/(\alpha + \beta + \gamma)$. There does not appear to be a convincing biological rationale for this model with $\beta \neq \gamma$. However, the extra parameter allows some more flexibility in fitting to data. Moreover, the analysis of the 3 parameter model is no more difficult than that of the 2 parameter one, and is even somewhat clearer from an expository point of view. We refer the reader to [ES93, EZ98] for the changes that are necessary in what follows when dealing with the 1 and 2 parameter models.

Probabilists usually record the rates for a Markov chain as an *infinitesimal generator matrix*. For example, the infinitesimal generator for the 3 parameter Kimura model is

$$Q = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \quad A \quad & \quad G \quad & \quad C \quad & \quad T \quad \end{array} \\ \begin{pmatrix} -(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & -(\alpha+\beta+\gamma) & \gamma & \beta \\ \beta & \gamma & -(\alpha+\beta+\gamma) & \alpha \\ \gamma & \beta & \alpha & -(\alpha+\beta+\gamma) \end{pmatrix} \end{array}.$$

The infinitesimal generator is more than just an accounting device: for any $s, t \geq 0$ the entry in row B' and column B'' of the matrix

$$\exp(tQ) = I + tQ + \frac{t^2}{2!}Q^2 + \frac{t^3}{3!}Q^3 + \cdots$$

gives the conditional probability that nucleotide B'' will be exhibited at time $s+t$ given that nucleotide B' is exhibited at time s.

Because the matrix Q is symmetric, $\exp(tQ)$ can be computed using the spectral theorem once the eigenvalues and eigenvectors of Q have been computed. This is straightforward for Q, but we won't go into the details. Also, the diagonalisation follows easily using the Fourier ideas of Section 6. As an example, the conditional probability that nucleotide A will be exhibited at time $s+t$ given that nucleotide A is exhibited at time s is

$$\frac{1}{4}[1+\exp(-2t(\alpha+\gamma))+\exp(-2t(\beta+\gamma))+\exp(-2t(\alpha+\beta))],$$

and the the conditional probability that nucleotide G will be exhibited at time $s+t$ given that nucleotide A is exhibited at time s is

$$\frac{1}{4}[1-\exp(-2t(\alpha+\gamma))+\exp(-2t(\beta+\gamma))-\exp(-2t(\alpha+\beta))].$$

Note that both of these probabilities converge to $\frac{1}{4}$ as $t\to\infty$: of course, we expect from the symmetries of the Markov chain that if it evolves for a long time, then it will converge towards an equilibrium distribution in which all nucleotides are equally likely to be exhibited.

It is clear without computing $\exp(tQ)$ explicitly that this matrix is of the form

$$\begin{array}{c c} & \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{array}{c} A \\ G \\ C \\ T \end{array} & \begin{pmatrix} w & x & y & z \\ x & w & z & y \\ y & z & w & x \\ z & y & x & w \end{pmatrix} \end{array},$$

where $0\le w,x,y,z\le 1$. Not all such matrices are given by $\exp(tQ)$ for a suitable choice of α,β,γ,t. However, we suppose **from now on** that each substitution matrix $P^{(v)}$ is of this somewhat more general form for some w,x,y,z (that can vary with v.) Thus, once a tree $\mathbf{T}$ with m leaves and n vertices is fixed, there are $3n$ independent parameters in the model: 3 for the root distribution π and 3 for each of the $n-1$ substitution matrices. Note that each of the 4^m model probabilities $p((B_\ell)_{\ell\in\mathbf{L}})$, $(B_\ell)_{\ell\in\mathbf{L}}\in\{A,G,C,T\}^{\mathbf{L}}$ is a polynomial in these $3n$ variables.

4. Making inferences

From the development in Sections 2 and 3, we have a model for the joint probability of the taxa exhibiting a particular set of nucleotides. For more than a small number of taxa, this model still has too many parameters for us to apply maximum likelihood. Moreover, maximum likelihood necessarily estimates all the numerical parameters in the model, even though the tree parameter is typically the one that is of most interest.

An alternative approach to estimating the tree that does not involve directly estimating the numerical parameters was suggested in [CF87] and [Lak87]. The ideas behind this approach is as follows. For a given tree $\mathbf{T}$, the model probabilities $p((B_\ell)_{\ell\in\mathbf{L}})$, $(B_\ell)_{\ell\in\mathbf{L}}\in\{A,G,C,T\}^{\mathbf{L}}$, have a specific functional form in terms of the numerical parameters defining the root distribution and the substitution matrices (indeed, the model probabilities are polynomials in these variables.) This should constrain the model probabilities to lie on some lower dimensional surface in $\mathbb{R}^{\mathbf{L}}$. Rather than represent this surface *explicitly* as the range of a vector of polynomials, we could try to characterise the surface *implicitly* as a subset of a locus of points in $\mathbb{R}^{\mathbf{L}}$ that are common zeroes of a family of polynomials. That is, we want to represent the surface as a subset of an *algebraic variety*.

Because we assuming that the same model (with the same numerical substitution mechanism parameters) governs each position in our data set and that the behaviour at different positions is independent, the strong law of large numbers gives that the quantities $p((B_\ell)_{\ell\in\mathbf{L}})$, $(B_\ell)_{\ell\in\mathbf{L}} \in \{A, G, C, T\}^{\mathbf{L}}$, can be consistently estimated in a model-free way by computing the proportion of positions in our data set at which Taxon 1 exhibits nucleotide B_1, Taxon 2 exhibits nucleotide B_2, *etc.* Call these estimates $\hat{p}((B_\ell)_{\ell\in\mathbf{L}})$, $(B_\ell)_{\ell\in\mathbf{L}} \in \{A, G, C, T\}^{\mathbf{L}}$, so that $\hat{p}((B_\ell)_{\ell\in\mathbf{L}})$ will be close to $p((B_\ell)_{\ell\in\mathbf{L}})$ with high probability when we observe a sufficient number of different positions to have enough independent identically distributed data points for the strong law of large numbers to kick in.

We hope that the varieties for two different trees (say, Tree I and Tree II) have a "small" intersection and so a "generic" point on the variety for one tree will not be a common zero of the polynomials defining the variety for the other tree. That is, we hope that we can find a polynomial f such that $f(p((B_\ell)_{\ell\in\mathbf{L}})) = 0$ for all choices of substitution mechanism parameters for Tree I whereas $f(p((B_\ell)_{\ell\in\mathbf{L}})) \neq 0$ for all but a "small" set of choices of substitution mechanism parameters for Tree II. If this is the case, then $f(\hat{p}((B_\ell)_{\ell\in\mathbf{L}}))$ should be close to zero (that is, "zero up to random error") if Tree I is the correct tree regardless of the numerical parameters in the model, whereas this quantity should be "significantly non-zero" if Tree II is the correct tree unless we have been particularly unfortunate and the numerical parameters are such that the vector $p((B_\ell)_{\ell\in\mathbf{L}})$ happens to lie on the intersection of the varieties for the two trees.

The polynomials that are zero on the algebraic variety associated with a tree are called the *(phylogenetic) invariants* of the model. Note that the set of invariants has the structure of an *ideal* in the ring of polynomials in the model probabilities: the sum of two invariants is an invariant and the product of an invariant with an arbitrary polynomial is an invariant.

In order to use the invariant idea to reconstruct phylogenetic trees we need to address the following questions:

i) How do we recognize when a polynomial is an invariant?
ii) How do we find a generating set for the ideal of invariants (and how big is such a set)?
iii) Do different trees have different invariants?
iv) How do we determine whether a vector of polynomials applied to estimates of the model probabilities is "zero up to random error" or "significantly non-zero"?

In principle, questions (i) and (ii) can be answered using general theory from computational commutative algebra. There is an algorithm using Gröbner bases that solves the *implicitization problem* of finding a generating set for the ideal of polynomials that are 0 on a general parametrically given algebraic variety (see [CLO92].) Unfortunately, this algorithm appears to be computationally infeasible for the size of problem that occurs for even a modest number of taxa. Other methods adapted to our particular problem are therefore necessary, and this is what we study in these notes. Along the way, we answer question (iii) and even establish how many algebraically independent invariants there are that distinguish between two trees. We don't deal with the more statistical question (iv) in these notes.

5. Some group structure

We begin with a step that may seem somewhat bizarre at first, but pays off handsomely. Consider the *Klein 4-group* $\mathbb{Z}_2 \bigoplus \mathbb{Z}_2$ consisting of the elements $\{(0,0),(0,1),(1,0),(1,1)\}$ equipped with the group operation of coordinatewise addition modulo 2. The addition table for $\mathbb{Z}_2 \bigoplus \mathbb{Z}_2$ is thus

$$\begin{array}{cc} + & \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{array}{c} (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} & \begin{pmatrix} (0,0) & (0,1) & (1,0) & (1,1) \\ (0,1) & (0,0) & (1,1) & (1,0) \\ (1,0) & (1,1) & (0,0) & (0,1) \\ (1,1) & (1,0) & (0,1) & (0,0) \end{pmatrix} \end{array}.$$

Identify the nucleotides $\{A,G,C,T\}$ with the elements of $\mathbb{Z}_2 \bigoplus \mathbb{Z}_2$ as follows: $A \leftrightarrow (0,0)$, $G \leftrightarrow (0,1)$, $C \leftrightarrow (1,0)$, and $T \leftrightarrow (1,1)$. This turns $\mathbb{G} := \{A,G,C,T\}$ into a group with the addition table

$$\begin{array}{cc} + & \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{array}{c} A \\ G \\ C \\ T \end{array} & \begin{pmatrix} A & G & C & T \\ G & A & T & C \\ C & T & A & G \\ T & C & G & A \end{pmatrix} \end{array}.$$

Suppose that X and Y are two $\mathbb{G}$-valued random variables such that the conditional distribution of Y given X is described by the matrix

$$\begin{array}{cc} & \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{array}{c} A \\ G \\ C \\ T \end{array} & \begin{pmatrix} w & x & y & z \\ x & w & z & y \\ y & z & w & x \\ z & y & x & w \end{pmatrix} \end{array}.$$

Note that $\mathbb{P}\{Y = B'' \,|\, X = B'\}$ only depends on the pair of nucleotides (B', B'') through the difference $B'' - B'$. It follows easily from this that the joint distribution of the pair (X, Y) is same as that of the pair $(X, X + Z)$, where $\mathbb{P}\{Z = A\} = w$, $\mathbb{P}\{Z = G\} = x$, $\mathbb{P}\{Z = C\} = y$, $\mathbb{P}\{Z = T\} = z$, and Z is independent of X.

The model that we described in Section 3 had an arbitrary root distribution π and substitution matrices $P^{(v)}$ that satisfy $P^{(v)}(B', B'') = q^{(v)}(B'' - B')$ for some probability distribution $q^{(v)}$ on $\mathbb{G}$. Repeatedly applying the observation of the previous paragraph shows that if if $(Z_v)_{v \in \mathbf{V}}$ is a vector of independent $\mathbb{G}$-valued random variables, with Z_ρ having distribution π, and Z_v, $v \in \mathbf{V} \backslash \{\rho\}$, having distribution $q^{(v)}$, then the $\mathbb{G}$-valued random variables

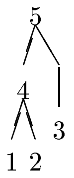
$$Y_\ell := \sum_{v \leq \ell} Z_v, \ \ell \in \mathbf{L},$$

have joint distribution

$$\mathbb{P}\{Y_1 = B_1, \ldots, Y_m = B_m\} = p((B_\ell)_{\ell \in \mathbf{L}}).$$

That is, by suitable addition of independent $\mathbb{G}$-valued "weights," we can construct a vector of random variables having the same joint distribution as the nucleotides exhibited by the taxa.

For example, for the tree



the construction is

$$\begin{array}{lllllllll} Y_1 & = & Z_1 & & & + & Z_4 & + & Z_5 \\ Y_2 & = & & Z_2 & & + & Z_4 & + & Z_5 \\ Y_3 & = & & & Z_3 & & & + & Z_5 \end{array}$$

6. A little Fourier analysis

We've seen that the model of Section 3 can be represented in terms of sums of indpendent random variables taking values in a finite, Abelian group. Probabilists have known for a long time that Fourier analysis is a very powerful technique for handling such sums. In this section we'll review some basic facts about Fourier analysis for an arbitrary finite, Abelian group $(\mathbb{H}, +)$.

Let $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ denote the unit circle in the complex plane, and regard $\mathbb{T}$ as an Abelian group with the group operation being ordinary complex multiplication. The *characters* of $\mathbb{H}$ are the group homomorphisms mapping $\mathbb{H}$ into $\mathbb{T}$. That is, $\chi : \mathbb{H} \to \mathbb{T}$ is a character if $\chi(h_1 + h_2) = \chi(h_1)\chi(h_2)$ for all $h_1, h_2 \in \mathbb{G}$. The characters form an Abelian group under the operation of pointwise multiplication of functions. This group is called the *dual group* of $\mathbb{H}$ and is denoted by $\hat{\mathbb{H}}$. The groups $\mathbb{H}$ and $\hat{\mathbb{H}}$ are isomorphic. Given $h \in \mathbb{H}$ and $\chi \in \hat{\mathbb{H}}$, write $\langle h, \chi \rangle$ for $\chi(h)$.

The elements of $\mathbb{H}$ form an orthogonal basis for the space of functions from $\mathbb{H}$ to $\mathbb{C}$. Given a function $f : \mathbb{H} \to \mathbb{C}$, the *Fourier transform* of f is the function $\hat{f} : \hat{\mathbb{H}} \to \mathbb{C}$ given by

$$\hat{f}(\chi) = \sum_{h \in \mathbb{H}} f(h) \langle h, \chi \rangle.$$

A function can be recovered from its Fourier transform via *Fourier inversion*:

$$f(h) = \frac{1}{\#\mathbb{H}} \sum_{\chi \in \hat{\mathbb{H}}} \hat{f}(\chi) \overline{\langle h, \chi \rangle}.$$

Given two finite, Abelian groups $\mathbb{H}'$ and $\mathbb{H}''$, the dual of the product group $\mathbb{H}'' \bigoplus \mathbb{H}''$ is isomorphic to $\widehat{\mathbb{H}'} \bigoplus \widehat{\mathbb{H}''}$ via the identification

$$\langle (h', h''), (\chi', \chi'') \rangle = \langle h', \chi' \rangle \times \langle h'', \chi'' \rangle.$$

One may write $\hat{\mathbb{G}} = \{1, \phi, \psi, \phi\psi\}$, where the following table gives the values of $\langle g, \chi \rangle$ for $g \in \mathbb{G}$ and $\chi \in \hat{\mathbb{G}}$:

$$\begin{array}{c c}
 & \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\
\begin{array}{c} 1 \\ \phi \\ \psi \\ \phi\psi \end{array} & \left(\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{array} \right).
\end{array}$$

The *characteristic function* of a $\mathbb{H}$-valued random variable X is the Fourier transform of its probability mass function:

$$\begin{aligned}
\xi(\chi) &= \sum_{h \in \mathbb{H}} \mathbb{P}\{X = h\} \langle h, \chi \rangle \\
&= \mathbb{E}[\langle X, \chi \rangle]
\end{aligned}$$

(here, following the usual convention in probability theory, $\langle X, \chi \rangle$ is the random variable obtained by composing the random variable X with the function $\langle \cdot, \chi \rangle$.) The probability mass function of X can be recovered from its Fourier transform by Fourier inversion:

$$\mathbb{P}\{X = h\} = \frac{1}{\#\mathbb{H}} \sum_{\chi \in \hat{\mathbb{H}}} \xi(\chi) \overline{\langle h, \chi \rangle}.$$

Finally, note that if X' and X'' are independent $\mathbb{H}$-valued random variables, then

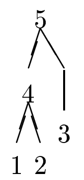
$$\mathbb{E}[\langle X' + X'', \chi \rangle] = \mathbb{E}[\langle X', \chi \rangle \langle X'', \chi \rangle] = \mathbb{E}[\langle X', \chi \rangle] \, \mathbb{E}[\langle X'', \chi \rangle].$$

That is, the characteristic function of $X' + X''$ is the product of the characteristic functions of X' and X''.

7. Finding an invariant

Let's begin by seeing how the observations of Sections 5 and 6 can be used to find an invariant for an instance of the model of Section 3.

Consider the tree



with the associated model for the nucleotides Y_1, Y_2, Y_3 exhibited by the taxa written in terms of independent $\mathbb{G}$-valued random variables $Z_1, \ldots, Z_5$ as follows:

$$\begin{array}{ccccccccc}
Y_1 & = & Z_1 & & & + & Z_4 & + & Z_5 \\
Y_2 & = & & Z_2 & & + & Z_4 & + & Z_5 \\
Y_3 & = & & & Z_3 & & & + & Z_5
\end{array}$$

Using the results of Section 6 and the notation given there for for the characters of $\mathbb{G}$ we have

$$\begin{aligned}
&\mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \phi\rangle\langle Y_3, \psi\rangle] \\
&= \mathbb{E}[\langle Z_1, \phi\rangle\langle Z_4, \phi\rangle\langle Z_5, \phi\rangle\langle Z_2, \phi\rangle\langle Z_4, \phi\rangle\langle Z_5, \phi\rangle\langle Z_3, \psi\rangle\langle Z_5, \psi\rangle] \\
&= \mathbb{E}[\langle Z_1, \phi\rangle]\,\mathbb{E}[\langle Z_2, \phi\rangle]\,\mathbb{E}[\langle Z_3, \psi\rangle]\,\mathbb{E}[\langle Z_4, \phi^2\rangle]\,\mathbb{E}[\langle Z_5, \phi^2\psi\rangle] \\
&= \mathbb{E}[\langle Z_1, \phi\rangle]\,\mathbb{E}[\langle Z_2, \phi\rangle]\,\mathbb{E}[\langle Z_3, \psi\rangle]\,\mathbb{E}[\langle Z_5, \psi\rangle].
\end{aligned}$$

A similar argument shows that

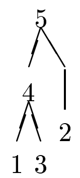
$$\begin{aligned}
&\mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \phi\rangle]\,\mathbb{E}[\langle Y_3, \psi\rangle] \\
&= \mathbb{E}[\langle Z_1, \phi\rangle]\,\mathbb{E}[\langle Z_2, \phi\rangle]\,\mathbb{E}[\langle Z_3, \psi\rangle]\,\mathbb{E}[\langle Z_5, \psi\rangle].
\end{aligned}$$

Thus

$$\mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \phi\rangle\langle Y_3, \psi\rangle] - \mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \phi\rangle]\,\mathbb{E}[\langle Y_3, \psi\rangle] = 0.$$

Writing all of the expectations in the last equation as sums in terms of the model probabilities $p((B_\ell)_{\ell\in\mathbf{L}})$ gives a polynomial in the model probabilities of total degree 2 that is satisfied for all choices of the numerical parameters defining the root distribution and the substitution matrices. Thus we have found an invariant for this tree.

Now consider the tree



with the associated model for the nucleotides Y_1, Y_2, Y_3 exhibited by the taxa written in terms of independent $\mathbb{G}$-valued random variables $Z_1, \ldots, Z_5$ as follows:

$$\begin{array}{ccccccccc}
Y_1 & = & Z_1 & & & + & Z_4 & + & Z_5 \\
Y_2 & = & & Z_2 & & + & & + & Z_5 \\
Y_3 & = & & & Z_3 & & Z_4 & + & Z_5
\end{array}$$

Now

$$\begin{aligned}
&\mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \phi\rangle\langle Y_3, \psi\rangle] - \mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \phi\rangle]\,\mathbb{E}[\langle Y_3, \psi\rangle] \\
&= \mathbb{E}[\langle Z_1, \phi\rangle]\,\mathbb{E}[\langle Z_2, \phi\rangle]\,\mathbb{E}[\langle Z_3, \psi\rangle]\,\mathbb{E}[\langle Z_4, \phi\psi\rangle]\,\mathbb{E}[\langle Z_5, \psi\rangle] \\
&\quad - \mathbb{E}[\langle Z_1, \phi\rangle]\,\mathbb{E}[\langle Z_2, \phi\rangle]\,\mathbb{E}[\langle Z_3, \psi\rangle]\,\mathbb{E}[\langle Z_4, \phi\rangle]\,\mathbb{E}[\langle Z_4, \psi\rangle]\,\mathbb{E}[\langle Z_5, \psi\rangle] \\
&= \mathbb{E}[\langle Z_1, \phi\rangle]\,\mathbb{E}[\langle Z_2, \phi\rangle]\,\mathbb{E}[\langle Z_3, \psi\rangle]\Big(\mathbb{E}[\langle Z_4, \phi\psi\rangle] - \mathbb{E}[\langle Z_4, \phi\rangle]\,\mathbb{E}[\langle Z_4, \psi\rangle]\Big)\mathbb{E}[\langle Z_5, \psi\rangle].
\end{aligned}$$

It is not hard to show that that the vector

$$(\mathbb{E}[\langle Z_4, \phi\rangle], \mathbb{E}[\langle Z_4, \psi\rangle], \mathbb{E}[\langle Z_4, \phi\psi\rangle])$$

ranges over a subset of $\mathbb{R}^3$ with non-empty interior as the distribution of Z_4 ranges over the set of possible distributions on $\mathbb{G}$. Thus

$$\mathbb{E}[\langle Z_4, \phi\psi\rangle] - \mathbb{E}[\langle Z_4, \phi\rangle]\,\mathbb{E}[\langle Z_4, \psi\rangle]$$

is certainly not identically 0 and the invariant we found for the previous tree is not an invariant for this tree.

8. Finding all invariants

The examples studied in Section 7 indicate how we should proceed to find all the invariants for a general tree. The ideas that we describe in this section were developed in [ES93].

We call a vector $(\chi_{\ell_1}, \ldots, \chi_{\ell_m}) \in \hat{\mathbb{G}}^m$ an *allocation of characters to leaves*. Such an allocation of characters to leaves induces an *allocation of characters to vertices* $(\chi_{v_1}, \ldots, \chi_{v_n}) \in \hat{\mathbb{G}}^n$ as follows. The character χ_v is the product of the χ_ℓ for all leaves ℓ that are descendents of v, that is,

$$\chi_v := \prod_{\ell \ge v} \chi_\ell.$$

In particular, if $v = v_i$ is a leaf (and hence the leaf ℓ_i by our numbering convention), then $\chi_{v_i} = \chi_{\ell_i}$.

Let

$$\{(\chi_{i,1}, \ldots, \chi_{i,n}),\, i = 1, \ldots, 4^m\}$$

be an enumeration of the various allocations of characters to vertices induced by the 4^m different allocations of characters to leaves. Define $3n$ vectors $\{\mathbf{x}_{v,\theta} = (x_{v,\theta}^{(1)}, \ldots, x_{v,\theta}^{(4^m)}),\, v \in \mathbf{V},\, \theta = \phi, \psi, \phi\psi\}$ of dimension 4^m by setting

$$x_{v_j,\theta}^{(i)} := \begin{cases} 1, \text{ if } \chi_{i,j} = \theta, \\ 0, \text{ otherwise}, \end{cases}$$

for $i = 1, \ldots, 4^m$, $j = 1, \ldots, n$ and $\theta \in \{\phi, \psi, \phi\psi\}$.

Write $\mathcal{R}(\mathbf{T})$ for the free $\mathbb{Z}$–module generated by the set $\{\mathbf{x}_{v,\theta} : v \in \mathbf{V},\, \theta = \phi, \psi, \phi\psi\}$. That is, $\mathcal{R}(\mathbf{T})$ is the collection of integer vectors of dimension 4^m consisting of $\mathbb{Z}$-linear combinations of the $\mathbf{x}_{v,\theta}$. Set

$$\mathcal{N}(\mathbf{T}) := \{a \in \mathbb{Z}^{4^m} : \sum_{i=1}^{4^m} a_i x_{v,\theta}^{(i)} = 0,\, v \in \mathbf{V},\, \theta = \phi, \psi, \phi\psi\},$$

so that $\mathbb{Z}^{4^m} = \mathcal{R}(\mathbf{T}) \oplus \mathcal{N}(\mathbf{T})$.

For $a \in \mathbb{Z}^{4^m}$, the polynomial

$$\begin{aligned} &\prod_{\{i : a_i \ge 0\}} \left(\mathbb{E}\left[\prod_{j=1}^m \langle Y_j, \chi_{i,j} \rangle \right] \right)^{a_i} - \prod_{\{i : a_i \le 0\}} \left(\mathbb{E}\left[\prod_{j=1}^m \langle Y_j, \chi_{i,j} \rangle \right] \right)^{-a_i} \\ &= \prod_{\{i : a_i \ge 0\}} \left(\sum_{(B_1, \ldots, B_m) \in \mathbb{G}^m} \prod_{j=1}^m \langle B_j, \chi_{i,j} \rangle p(B_1, \ldots, B_m) \right)^{a_i} \\ &\quad - \prod_{\{i : a_i \le 0\}} \left(\sum_{(B_1, \ldots, B_m) \in \mathbb{G}^m} \prod_{j=1}^m \langle B_j, \chi_{i,j} \rangle p(B_1, \ldots, B_m) \right)^{-a_i} \end{aligned}$$

is an invariant if and only if $a \in \mathcal{N}(\mathbf{T})$. It is shown in [ES93] that this is the only game in town: the invariants produced this way generate the ideal of all invariants.

Indeed, it is shown in [ES93] that if $\{(a_{1,r}, ..., a_{4^m,r}),\ r = 1, ..., \operatorname{rank} \mathcal{N}(\mathbf{T})\}$ is a $\mathbb{Z}$-basis for the free $\mathbb{Z}$-module $\mathcal{N}(\mathbf{T})$, then the set of polynomials of the form

$$\prod_{\{i: a_{i,r} \geq 0\}} \left(\mathbb{E}\left[\prod_{j=1}^{m} \langle Y_j, \chi_{i,j} \rangle \right] \right)^{a_{i,r}} - \prod_{\{i: a_{i,r} \leq 0\}} \left(\mathbb{E}\left[\prod_{j=1}^{m} \langle Y_j, \chi_{i,j} \rangle \right] \right)^{-a_{i,r}}$$

generates the ideal of invariants but no subset thereof does. Finding a $\mathbb{Z}$-basis for $\mathcal{N}(\mathbf{T})$ is just elementary linear algebra – we are simply finding a basis for the null space of an integer-valued matrix – and can be done using Gaussian elimination.

9. How many invariants are there?

Given our tree $\mathbf{T}$ with m leaves (taxa) and n vertices in total, we have 4^m model probabilities $p((B_\ell)_{\ell \in \mathbf{L}})$ that arise as polynomials in $3n$ "free parameters" — 3 free parameters for the root distribution and 3 free parameters for each of the substitution matrices. A naive "degrees of freedom" argument would suggest that there should, in some sense, be $4^m - 3n$ independent relations between the model probabilities. We verify this numerology in this section by showing that $\operatorname{rank} \mathcal{R}(\mathbf{T}) = 3n$, and hence $\operatorname{rank} \mathcal{N}(\mathbf{T}) = 4^m - 3n$. This and related results were presented in [EZ98], but our proof here is quite different.

Let $\mathbf{X}$ denote the $4^m \times 3n$ matrix with columns indexed by $\mathbf{V} \times \{\phi, \psi, \phi\psi\}$ that has the column corresponding to (v, θ), given by $\mathbf{x}_{v,\theta}$. We need to show that the matrix $\mathbf{X}$ has (real) rank $3n$, and this is equivalent to showing that the associated $3n \times 3n$ Gram matrix $\mathbf{X}^t\mathbf{X}$ has full rank (see 0.4.6(d) of [HJ85].)

The entry of $\mathbf{X}^t\mathbf{X}$ with indices $((v^*, \theta^*), (v^{**}, \theta^{**}))$, $v^*, v^{**} \in \mathbf{V}$, $\theta^*, \theta^{**} \in \{\phi, \psi, \phi\psi\}$, is the usual scalar product of $\mathbf{x}_{v^*,\theta^*}$ with $\mathbf{x}_{v^{**},\theta^{**}}$, which is just the number of assignments of characters to leaves that assign θ^* to v^* and θ^{**} to v^{**}. We can compute this number of assignments as follows.

If $v^* = v^{**}$ and $\theta^* = \theta^{**}$, then it is clear by symmetry that this entry is 4^{m-1}, whereas if $v^* = v^{**}$ and $\theta^* \neq \theta^{**}$, then this entry is obviously 0.

Consider now the case where $v^* \neq v^{**}$, so that the collection of leaves descended from v^* is not the same as the collection of leaves descended from v^{**}. We claim that the entry of $\mathbf{X}^t\mathbf{X}$ with indices $((v^*, \theta^*), (v^{**}, \theta^{**}))$ is 4^{m-2}. To see this, write $\mathbf{L}^*$ and $\mathbf{L}^{**}$ for the leaves descended from v^* and v^{**}, respectively. Suppose first that $\mathbf{L}^{**} \subsetneq \mathbf{L}^*$. If we have an assignment of characters to leaves that assigns the characters η^* to v^* and η^{**} to v^{**}, then replacing the character assigned to some $\ell^* \in \mathbf{L}^* \backslash \mathbf{L}^{**}$ from χ^* (say) to $\rho^*\rho^{**}\eta^*\chi^*$ and replacing the character assigned to some $\ell^{**} \in \mathbf{L}^{**}$ from χ^{**} (say) to $\rho^{**}\eta^{**}\chi^{**}$ gives a new assignment of characters to leaves that assigns ρ^* to v^* and ρ^{**} to v^{**}. It follows that number of assignments of characters to leaves that assign θ^* to v^* and θ^{**} to v^{**} is indeed 4^{m-2} when $\mathbf{L}^{**} \subsetneq \mathbf{L}^*$. A symmetric argument argument handles the case $\mathbf{L}^* \subsetneq \mathbf{L}^{**}$, and we leave this to the reader.

We conclude that $\mathbf{X}^t\mathbf{X}$ can be partitioned into 3×3 blocks so that the blocks down the diagonal are all of the form

$$\begin{pmatrix} 4^{m-1} & 0 & 0 \\ 0 & 4^{m-1} & 0 \\ 0 & 0 & 4^{m-1} \end{pmatrix},$$

while the off–diagonal blocks are all of the form

$$\begin{pmatrix} 4^{m-2} & 4^{m-2} & 4^{m-2} \\ 4^{m-2} & 4^{m-2} & 4^{m-2} \\ 4^{m-2} & 4^{m-2} & 4^{m-2} \end{pmatrix}.$$

Now

$$\mathbf{X}^t \mathbf{X} = 4^{m-2}(\mathbf{D} + \mathbf{1}\mathbf{1}^t)$$

where $\mathbf{1}$ is the (column) vector with all entries equal to 1 and $\mathbf{D}$ is a matrix partitioned into 3×3 blocks with the blocks down the diagonal all of the form

$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix},$$

and the off–diagonal blocks all zero. Note that $\mathbf{D}$ is invertible with inverse a partitioned matrix that has blocks down the diagonal all of the form

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix},$$

and the off–diagonal blocks all zero. A standard result on inverses of small rank perturbations (see 0.7.4 of [HJ85]) gives that $\mathbf{X}^t\mathbf{X}$ is indeed invertible (and hence full rank), with inverse

$$\begin{aligned} & 4^{-(m-2)} \left(\mathbf{D}^{-1} - \frac{1}{1 + \mathbf{1}^t \mathbf{D}^{-1} \mathbf{1}} \mathbf{D}^{-1} \mathbf{1}\mathbf{1}^t \mathbf{D}^{-1} \right) \\ & = 4^{-(m-2)} \left(\mathbf{D}^{-1} - \frac{1}{1 + 3n} \mathbf{1}\mathbf{1}^t \right). \end{aligned}$$

10. How well do invariants distinguish between trees?

The last question remaining from Section 4 is, “Do different trees have different invariants?” The answer is “Yes.” This follows from Theorem 10 in [SSE93]. We give a different proof which actually establishes “how many” independent invariants distinguish between two different trees.

We begin by making explicit the natural notion of equivalence for trees with labelled leaves. We say that two trees $\mathbf{T}'$ and $\mathbf{T}''$ with the same set $\mathbf{L}$ of leaves are *identical* if there is a bijection τ from the set of vertices $\mathbf{V}'$ of $\mathbf{T}'$ to the set of vertices $\mathbf{V}''$ of $\mathbf{T}''$ such that $\tau(\ell) = \ell$ for each leaf $\ell \in \mathbf{L}$ and $u \in \mathbf{V}'$ is the father of $v \in \mathbf{V}'$ in $\mathbf{T}'$ if and only if $\tau(u) \in \mathbf{V}''$ is the father of $\tau(v) \in \mathbf{V}''$ in $\mathbf{T}''$. This is equivalent to requiring that $\tau(\ell) = \ell$ for each leaf $\ell \in \mathbf{L}$ and $u \in \mathbf{V}'$ is the ancestor of $v \in \mathbf{V}'$ in $\mathbf{T}'$ if and only if $\tau(u) \in \mathbf{V}''$ is the ancestor of $\tau(v) \in \mathbf{V}''$ in $\mathbf{T}''$. It is not hard to see that two trees $\mathbf{T}'$ and $\mathbf{T}''$ with the same set $\mathbf{L}$ of leaves are identical if and only if for each $v' \in \mathbf{V}'$ the set of leaves descended from v' is equal to the set of leaves descended from some $v'' \in \mathbf{V}''$ and vice-versa.

Given two trees $\mathbf{T}'$ and $\mathbf{T}''$ with the same set $\mathbf{L}$ of leaves, write $\nu(\mathbf{T}', \mathbf{T}'')$ for the number of vertices v'' of $\mathbf{T}''$ such that the collection of leaves descended from v'' is not the collection of leaves descended from any vertex of $\mathbf{T}'$. If $\mathbf{T}'$ and $\mathbf{T}''$ are not identical, then either $\nu(\mathbf{T}', \mathbf{T}'') > 0$ or $\nu(\mathbf{T}'', \mathbf{T}') > 0$. We claim that the rank of the free $\mathbb{Z}$–module $\mathcal{N}(\mathbf{T}') \cap \mathcal{R}(\mathbf{T}'')$ is $3\nu(\mathbf{T}', \mathbf{T}'')$. That is, there are $3\nu(\mathbf{T}', \mathbf{T}'')$

algebraically independent invariants for the tree $\mathbf{T}'$ that are not invariants for the tree $\mathbf{T}''$, and similarly with the roles of $\mathbf{T}'$ and $\mathbf{T}''$ interchanged.

To establish this claim, first note that

$$\begin{aligned}\operatorname{rank}(\mathcal{N}(\mathbf{T}')\cap\mathcal{R}(\mathbf{T}'')) &= \operatorname{rank}(\mathcal{R}(\mathbf{T}'')) - \operatorname{rank}(\mathcal{R}(\mathbf{T}')\cap\mathcal{R}(\mathbf{T}'')) \\ &= \operatorname{rank}(\mathcal{R}(\mathbf{T}')+\mathcal{R}(\mathbf{T}'')) - \operatorname{rank}(\mathcal{R}(\mathbf{T}')).\end{aligned}$$

Write $\mathbf{V}'$ and $\mathbf{V}''$ for the vertices of $\mathbf{T}'$ and $\mathbf{T}''$, respectively, and let $\tilde{\mathbf{V}}''$ denote the set of vertices v'' of $\mathbf{T}''$ such that the collection of leaves descended from v'' is not the collection of leaves descended from any vertex of $\mathbf{T}'$. Hence $|\tilde{V}''| = \nu(\mathbf{T}', \mathbf{T}'')$. Of course, if $v'' \in \mathbf{V}''\backslash\tilde{\mathbf{V}}''$, then there is a vertex $v' \in \mathbf{V}'$ such that the assignment of characters to v' and v'' for each assignment of characters to leaves are the same, and hence the vector $\mathbf{x}_{v',\theta}$ (calculated for $\mathbf{T}'$) is the same as the vector $\mathbf{x}_{v'',\theta}$ (calculated for $\mathbf{T}''$.) The claim will thus follow if we can show that the vectors

$$\{\mathbf{x}_{v',\theta} : v' \in \mathbf{V}',\, \theta = \phi, \psi, \phi\psi\} \cup \{\mathbf{x}_{v'',\theta} : v'' \in \tilde{\mathbf{V}}'',\, \theta = \phi, \psi, \phi\psi\}$$

are linearly independent over the integers (equivalently, over the reals.)

Let $\mathbf{X}$ denote the $4^m \times 3(|\mathbf{V}'| + |\tilde{\mathbf{V}}''|)$ matrix with columns indexed by $(\mathbf{V}' \cup \tilde{\mathbf{V}}'') \times \{\phi, \psi, \phi\psi\}$ that has the column corresponding to (v', θ), $v' \in \mathbf{V}'$ (resp. (v'', θ), $v'' \in \tilde{\mathbf{V}}''$) given by $\mathbf{x}_{v',\theta}$ (resp. $\mathbf{x}_{v'',\theta}$.) We need to show that $\mathbf{X}$ has (real) rank $3(|\mathbf{V}'| + |\tilde{\mathbf{V}}''|)$, and this is equivalent to showing that the associated $3(|\mathbf{V}'|+|\tilde{\mathbf{V}}''|) \times 3(|\mathbf{V}'|+|\tilde{\mathbf{V}}''|)$ Gram matrix $\mathbf{X}^t\mathbf{X}$ has full rank. An argument very similar to that in Section 9 completes the proof.

References

[CF87] J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. Classification*, 4:57–71, 1987.

[CLO92] D. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms : an introduction to computational algebraic geometry and commutative algebra*. New York : Springer-Verlag, 1992.

[ES93] S.N. Evans and T.P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21:355–377, 1993.

[EZ98] S.N. Evans and X. Zhou. Constructing and counting phylogenetic invariants. *J. Comput. Biol.*, 5:713–724, 1998.

[GW91] Larry Gonick and Mark Wheelis. *The cartoon guide to genetics*. Harper Perennial, New York, updated edition, 1991.

[HJ85] R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1985.

[JC69] T.H. Jukes and C. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. New York: Academic Press, 1969.

[Kim80] M. Kimura. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.

[Kim81] M. Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, 78:454–458, 1981.

[Lak87] J.A. Lake. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Mol. Biol. Evol.*, 4:167–191, 1987.

[Ney71] J. Neyman. Molecular studies of evolution: A source of novel statistical problems. In S.S. Gupta and J. Yackel, editors, *Statistical Decision Theory and Related Topics*, pages 1–27. New York: Academic Press, 1971.

[SSE93] L. A. Székely, M. A. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. *Adv. in Appl. Math.*, 14(2):200–210, 1993.

[Wat95] Michael S. Waterman. *Introduction to computational biology : maps, sequences and genomes*. Chapman & Hall, London, New York, 1995.

E-mail address: evans@stat.Berkeley.EDU

Department of Statistics #3860, University of California at Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A