The Combinatorics of Evolutionary Trees—a Survey

L. A. Székely,* P. L. Erdős, M. A. Steel

ABSTRACT: We survey here results and problems from the reconstruction theory of evolutionary trees, which involve enumeration and inversion.

1. Introduction

Since the work of Darwin, there has been a dream of biologists: to reconstruct the tree of evolution of living things. That tree could be the *only* scientific basis for classification. In the last two decades the dramatic progress in molecular biology (reading long segments of genetic sequences) led to a new field, the theory of *molecular evolution*.

One assumes that the process of evolution is described by a tree, in which no degree exceeds 3, since evolutionary events are too rare to coincide. In this tree the leaves denote existing species represented by corresponding segments of aligned DNA sequences, the unlabelled branching vertices may denote unknown extinct ancestors; since fossils do not keep records of the DNA sequence. For a given set of existing species, we define their *true tree* by taking the subtree induced by them in the tree describing the process of evolution and undoing the vertices of degree two. We term any binary tree, in which leaves are labelled by the species and the branching vertices are unlabelled, an *evolutionary tree*. The very problem of reconstruction may be put in this way: given a set of species with corresponding segments of aligned DNA sequences, select the true tree from the set of possible evolutionary trees.

In this paper we assume that every bit of the aligned DNA sequence is one of the four nucleotides, A (Adenine), G (Guanine), C (Cytosine), T (Thymine); i.e. we neglect insertions and deletions. Biologists also would like to subdivide an edge of the true tree by a root r to denote a common ancestor and the direction of the evolution. However, if you have a procedure to solve the problem above, it easily can be applied to finding the root by *outgroup comparison*: add a new species to your list which is known to be far from all

* Research of the first author was supported by the A. v. Humboldt-Stiftung and the U. S. Office of Naval Research under the contract N-0014-91-J-1385

your species, reconstruct the larger true tree, and the neighbor of the new species can be considered the root of the smaller true tree.

It is not always the case, that A, G, C, T are the letters of the alphabet; a two-letter alphabet (identifying purines $A = G$ and pyrimidines $C = T$), and a 20-letter alphabet of amino acids for protein sequences are also possible.

To solve the reconstruction problem, one needs a mathematical model that distinguishes the true tree in mathematical terms, and one also may expect, that the mathematical model in question corresponds to a known or generally assumed mechanism of molecular evolution. One also may expect several other attributes of the model, as Hendy, Penny, and Steel [PHS1] listed: a polynomial time algorithm for tree reconstruction, convergence on relatively short sequences to the true tree, insensitivity to small errors in input data, and falsifiabilty of the model in a Popperian sense. However, no tree reconstruction method proposed is powerful enough to meet all these criteria; many popular ones do not even correspond to any assumed mechanism of molecular evolution. It is no surprise, that Penny, Hendy, Zimmer and Hamby [PHZH] can show sets of species, for which different evolutionary trees have been published on the basis of different data, and even on the basis of the same data, using different methods. In [PHS1], [PHS2], and other papers, Penny *et al.* gave a program to put the theory of evolutionary trees on a sound philosophical and mathematical foundation.

It is not the point of the present paper to overview advantages and shortcomings of all tree reconstruction methods. For a comparison of different methods, see [PHS1]. We restrict the present paper to our modest contribution, that involves enumeration and inversion, to that program. Sections 3-5 closely follow [SSE]. We give no proofs.

Cavalli-Sforza and Edwards [CSE] introduced the *parsimony principle* to the analogy of many minimum principles in science. In many instances the parsimony principle yields reasonably good trees, however no mechanism of evolution is accountable for it, and there are situations—where some branches of the true tree have significantly different rate of change—in which it may be false, see Felsenstein [F]. Section 2 is devoted to the parsimony principle and related enumeration results.

Section 3 describes a Fourier inverse pair depending on trees and Abelian groups, and specializes it to the group Z_2^m. Section 4 sets Kimura's models of molecular evolution in terms of Section 3 and outlines the spectral analysis/closest tree method. Section 5 is devoted to the construction of a complete set of invariants for Cavender's model and Kimura's 3-parameter model, and Section 6 concludes.

2. The parsimony principle

Let $\mathcal{C}$ denote the letters of our alphabet, which frequently will be referred to as a set of *colours*, and let $\mathcal{C}_m$ denote the set of $m-$letter words over that alphabet. Let T be an evolutionary tree with leaf set L. We term a map $\chi : L \longrightarrow \mathcal{C}_m$ as a *leaf-colouration.* The colouration $\bar{\chi} : V(T) \longrightarrow \mathcal{C}_m$ is an *extension* of the leaf-colouration χ if the two maps are identical on the set L. The *changing number* of the colouration $\bar{\chi}$ is the number of pairs of <edge, letter position>, where end-vertices of the edge have different colours at the corresponding letter position according to $\bar{\chi}$. We term the minimum changing number of the tree T over all extensions of χ the *length* of T. The *parsimony principle* says, that the true tree has minimum length, i.e. maximum parsimony. Unfortunately, results of Foulds and Graham [FG] show that the decision problem, whether for a set of leaves and assigned words, an evolutionary tree with prescribed length exists, is NP-hard, even when $|\mathcal{C}| = 2$. Therefore, from a statistical point of view, it is reasonable to ask for the expectation and variance of the length of a random evolutionary tree, in order to use this information as a selection principle (Steel [S1]). Not much is known yet on the variance, but there are some results on the expectation. The computation of the expectation can be reduced to the solution of the following enumeration problem.

Problem. *Let $f_k(a_1, ..., a_t)$ ($t \geq 2$, $a_i \geq 1$, $n = a_1 + \cdots + a_t$) denote the number of binary trees with a_i labelled leaves of colour i, with unlabelled branching vertices, with length k. Evaluate $f_k(a_1, ..., a_t)$.*

This enumeration problem is still open; not even a conjectured value of $f_k(a_1, ..., a_t)$ is at hand. We list here the solved instances of the problem. Carter, Hendy, Penny, Székely and Wormald [CHPSW] proved the

Bichromatic binary tree theorem.

$$f_k(a, b) = (k-1)!(2n-3k)N(a, k)N(b, k)\frac{(2n-5)!!}{(2n-2m-1)!!}, \tag{1}$$

where $a + b = n$ and

$$N(x, k) = \binom{2x-k-1}{k-1}(2x-2k-1)!!. \tag{2}$$

For more than 2 colours, results for extreme length values are available. Observe that with k colours present, the length is at least $k-1$. For this extreme value, Carter & al. [CHPSW] proved

$$f_{k-1}(a_1,...,a_k) = \frac{(2n-5)!!}{(2n-2k-1)!!} N(a_1,1)\cdots N(a_r,1).$$

For $a_i \geq 2$, using inclusion-exclusion, Steel [S1] went further to prove

$$f_k(a_1,...,a_k) = \frac{(k-1)(4(n-k)^2-2n+k)(2n-5)!!}{(2n-2k+1)!!} N(a_1,1)\cdots N(a_k,1).$$

In another paper Steel [S2] obtained:

$$f_{2k}(k,k,k) = (k!)^3 \sum_{s=1}^{k} [x^k] \frac{Q(x)^s}{s!} \frac{(6k-5)!!}{(6k-2s-1)!!}, \tag{3}$$

where $[x^i]Q(x) = \frac{2(4i-3)!(6i-3)}{(3i-1)!i!}$. Notice that with 3 colour classes of size k the length is at most $2k$, an extreme case, again. D. Penny [personal communication] computed some small values of f for 3 colours, which may be useful for making and/or checking conjectures:

$f_m(2,2,3) = 27,\ 318,\ 600$ for $m = 2,3,4$;

$f_m(2,2,4) = 165,\ 2610,\ 7620$ for $m = 2,3,4$;

$f_m(2,3,3) = 99,\ 1566,\ 5526,\ 3204$ for $m = 2,3,4,5$;

$f_m(3,3,3) = 351,\ 6966,\ 40554,\ 60858,\ 19116$ for $m = 2,3,4,5,6$;

$f_m(2,2,5) = 1365,\ 27090,\ 106680$ for $m = 2,3,4$;

$f_m(2,3,4) = 585,\ 11610,\ 57420,\ 65520$ for $m = 2,3,4,5$.

A trivial, but useful formula in establishing more values of f is

$$f_k(a_1,...,a_r,1) = (2n-5) f_{k-1}(a_1,...,a_r). \tag{4}$$

Using (1) and (4), one easily extends the little table above for the values of $f_m(1,a,b)$.

The first proof of the bichromatic binary tree theorem relied on generating functions, multivariate Lagrange inversion and computer algebra. Later on, Steel gave a proof from a combinatorial decomposition based on Menger's theorem [S1], and Erdős and Székely [ES2] simplified his proof further. It has turned out, that (2) counts binary forests of

k components on x labelled leaves, such that every component contains one vertex of degree two or zero [CHPSW], [E]. The term $k!N(a,k)N(b,k)$ nearly present in (1) can be explained as such forests being built on both colour classes of leaves and then the trees are matched in all possible ways. Then the rest of (1) comes into play at building different trees of length k from the matched forests.

It became evident, that a solution of the general enumeration problem requires a good characterization of the fact, that the length of a tree is not less than t; for two colours Menger's theorem provides for such a good characterization. A natural generalization of the length is the well-known *multiway cut* problem; given a graph G and $N \subseteq V(G)$, find an edge set of minimum size, whose deletion separates each pairs of N. Dalhaus & al. [DJPSY] showed that the multiway cut problem is NP-hard (even for planar graphs, if $|N|$ is not bounded). Hence, the existence of such a good characterization is unlikely in general. For $r \geq 2$ colours and (not necessarily binary) trees Erdős and Székely [ES3] proved the following min-max theorem to give good characterization:

Theorem. *The length of a leaf coloured tree is equal to the maximum number of oriented paths, connecting differently coloured leaves, such that no edge is used by two oppositely oriented paths, and no two paths using the same edge end in the same colour.*

However, this is not enough in itself, to solve the problem. Notice that it is unlikely that a product formula like (1) solves the problem, since the given numerical values have some large prime factors e.g. 43, 53, 89; and (3) does not suggest any closed form either.

We would like to close this section with applications and a by-product. The applications are in biology. The well-known astronomer Sir Fred Hoyle has suggested that the Earth is continually bombarded by viruses (including influenza viruses) that originate from comets. Henderson, Hendy and Penny [HHP] showed that his hypothesis may be rejected with very high probability; their basic mathematical tool was the bichromatic binary tree theorem. A further similar application, due to Steel, Hendy and Penny [SHP], applies the bichromatic binary tree theorem to calculate a permutation-based statistic for aligned sequences over the 2-letter alphabet, which allows for a test, whether the alignment is significantly "tree-like".

The byproduct is a bijection of Erdős and Székely [ES1] between some trees with unlabelled branching vertices and set partitions, which gives a unified technique to solve a number of tree enumeration problems. The motivation for the bijection came from counting evolutionary trees, which yields a semifactorial function (Cavalli-Sforza and Edwards [CSE]), like the number of partitions of a $2n$-element set into 2-element sets. Had not we seen counting of trees with unlabelled branching vertices in biomathematics, we would hardly have ever come to this point.

3. A Fourier calculus

Let us be given a tree T with leaf set L and one arbitrary leaf R, called a *root*. We do not need that the tree is binary, but we assume that no vertex has degree two. Suppose that we are given a finite Abelian group G and for the edges $e \in E(T)$ we have independent G-valued random variables ξ_e with $p_e(g) := Prob(\xi_e = g)$ and $\sum_{g \in G} p_e(g) = 1$. Produce a random G-colouration of the leaves of the tree by evaluating ξ_e for every edge and giving as colour to the leaf l the product of group elements along the unique Rl path. Let f_σ denote the probability that we obtain a leaf colouration $\sigma : \ L \setminus \{R\} \longrightarrow G$. We need to recall some facts on characters and the Fourier transform, which can be found in [J] or in [EvS].

Lemma. *Let G be a finite Abelian group, then*

(i) the character group $\hat{G}$ is isomorphic to G.

(ii) if $f : G \to C$ is a complex-valued function and $\hat{f} : \hat{G} \to C$ is defined by

$$\hat{f}(\chi) = \sum_{g \in G} \chi(g) f(g),$$

then for all $g \in G$

$$f(g) = \frac{1}{|G|} \sum_{\chi \in \hat{G}} \overline{\chi(g)} \hat{f}(\chi).$$

(iii) The characters of a direct product group are exactly the products of characters.

Take G^{n-1} = the set of colourations $\sigma : \ L \setminus \{R\} \longrightarrow G$ endowed with pointwise multiplication. Let $\chi = (\chi_l \in \hat{G} : l \in L \setminus \{R\})$ be an ordered $(n-1)$-tuple of characters. Then $\chi \in \hat{G}^{n-1}$ acts on G^{n-1} according to Lemma (iii). For $e \in E(T)$, set $L_e = \{l \in L : e \text{ separates } l \text{ from } R \text{ in } T\}$. Define

$$r_\chi = \prod_{e \in E(T)} \sum_{g \in G} p_e(g) \prod_{l \in L_e} \chi_l(g). \tag{5}$$

In [SSE] we obtained the following inverse pair:

Theorem.

$$r_\chi = \sum_{\sigma \in G^{n-1}} f_\sigma \prod_l \chi_l(\sigma(l)) \quad \textit{and} \tag{6}$$

$$f_\sigma = \frac{1}{|G|^{n-1}} \sum_{\chi \in \hat{G}^{n-1}} r_\chi \prod_l \overline{\chi_l(\sigma(l))}. \tag{7}$$

In [SSE] we observed that (6) and (7) are equivalent by Lemma (ii) for *any* $f : G^{n-1} \longrightarrow C$ and $r : \hat{G}^{n-1} \longrightarrow C$; and it is no longer difficult to prove (6) with *our* f_σ and r_χ, as soon as the correct definition (5) is discovered. Let us specialize the Theorem for $G = Z_2^m$, since it admits a combinatorial description and has practical significance at the same time. Fourier calculus over Z_2^m occured many times in the literature (see [SSE] for some references), but not for the same purpose.

For every $l \in L \setminus \{R\}$, take a copy of G, $G = G_l = Z_2^m$ and $\hat{G} = \hat{G}_l$ by

$$G_l = \{(\sigma_1, ..., \sigma_m) : \sigma_i \subseteq \{l\}\}, \quad \hat{G}_l = \{(X_1, ..., X_m) : X_i \subseteq \{l\}\},$$

both endowed with the positionwise symmetric difference operation as group multiplication. For $\chi_l = (X_1, ..., X_m) \in \hat{G}_l$ and $g_l = (\sigma_1, ..., \sigma_m) \in G_l$ define the action

$$\chi_l(g_l) = (-1)^{\sum_{i=1}^m |\sigma_i \cap X_i|}.$$

For the direct product of G_l's and $\hat{G}_l$'s one has

$$G^{n-1} = \{(\sigma_1, ..., \sigma_m) : \sigma_i \subseteq L \setminus \{R\}\}, \quad \hat{G}^{n-1} = \{(X_1', ..., X_m') : X_i' \subseteq L \setminus \{R\}\}.$$

For the combinatorial interpretation the key observation is that the latter formula can be identified with

$$\hat{G}^{n-1} = \{(X_1, ..., X_m) : X_i \subseteq L,\ |X_i| \text{ even}\},$$

endowed by the positionwise symmetric difference operation as group multiplication and character action

$$(X_1, ..., X_m)(\sigma_1, ..., \sigma_m) = (-1)^{\sum_{i=1}^m |\sigma_i \cap X_i|}$$

under the correspondence

$$X_i = \begin{cases} X_i', & \text{if } |X_i'| \text{ even}, \\ X_i' \cup \{R\}, & \text{if } |X_i'| \text{ odd}. \end{cases}$$

Now (5) turns into

$$r_{X_1,\dots,X_m} = \prod_{e\in E(T)} \sum_{g\subseteq M} (-1)^{|g\,\cap\,\{i\,:\,e\,\in\,P(T,X_i)\}|} p_e(g),$$

where $P(T,X_i) = \{e \in E(T) : |L_e \cap X_i| \text{ odd}\}$ (the unique *T-join* of the set X_i for a graph theorist); and (6)-(7) turns into (8)-(9)

$$r_{X_1,\dots,X_m} = \sum_{\sigma_1,\dots,\sigma_m} (-1)^{\sum_{j=1}^m |\sigma_j \cap X_j|} f_{\sigma_1,\dots,\sigma_m}, \tag{8}$$

$$f_{\sigma_1,\dots,\sigma_m} = \frac{1}{2^{m(n-1)}} \sum_{X_1,\dots,X_m} (-1)^{\sum_{j=1}^m |\sigma_j \cap X_j|} r_{X_1,\dots,X_m}. \tag{9}$$

It is an important fact that the connecting matrices in (6)-(7) are— after normalization— unitary, and hence the connecting matrices in (8)-(9) are Hadamard.

4. Kimura's models of molecular evolution

After the work of Kimura, the general assumption for the mechanism of molecular evolution is that changes in the DNA are *random*. It is assumed that changes at different sites are independent and of identical distribution. In case the data violates too much the condition on identical distribution, one may thin out the sequences by considering one site of each of the *codons* (the consecutive triplets of nucleotides encoding amino acids), particularly the third position, which is more redundant in the coding scheme than the other two positions, and therefore less influenced by natural selection. For $m = 1$, the model described in Section 3 specializes to a model of Cavender [C1], for which Hendy and Penny found the special case of the calculus above and applied it in their spectral analysis/closest tree method for tree reconstruction from sequences over a 2-letter alphabet [H], [HP1], [HP2]. Our part was the generalization for other groups; the practical importance of this generalization is mostly for $m = 2$, i.e. for sequences over the 4-letter alphabet A, G, C, T. We explain the $m = 2$ case in details, the explanation also applies, mutatis mutandis, to $m = 1$. It is an interesting paradox of the theory of evolution, that evolution is random at the molecular level and follows natural selection at a high level.

From now on we describe Kimura's 3-parameter model [K2, K3] and some restricted versions of it, which are known as Kimura's 2-parameter model [K1] and Jukes-Cantor model

[JC], (the Jukes-Cantor model is more explicit in Neyman [N]). We follow the group theoretical setting of these models from Evans and Speed [EvS]. Take the symmetric difference group of the subsets of $M = \{1, 2\}$, which is the Kleinian group $Z_2 \times Z_2$ with generators $\{1\}$ and $\{2\}$. *We compromise at this point and do not assume any longer, that evolutionary trees and the true tree are binary, but we assume, that they have no vertex of degree two.* Take the true tree with a common ancestor r, a random subset of M is assigned under a certain (unknown) distribution to r. To every edge of the tree a random element of the Kleinian group is assigned independently. In the group theoretical setting of the Kimura's models [EvS], the elements of the Kleinian group are identified with nucleotides, $A \leftrightarrow \emptyset$, $G \leftrightarrow \{2\}$, $C \leftrightarrow \{1\}$, $T \leftrightarrow M$. The random group element at r tells the original nucleotide value there, and the random variable at an edge describes the nucleotide change on that edge. In terms of biology, multiplication by $\emptyset$ on an edge causes no change in the nucleotide, multiplication by $\{2\}$ causes *transition*, and multiplication by $\{1\}$ or M causes one of the two possible types of *transversions.* To every leaf l the product of group elements along the unique path rl and in r itself is assigned. We have a random 4-colouration of the leaves (in fact, of all vertices) of the tree. That is Kimura's 3-parameter model of molecular evolution. Kimura's 3-parameter model allows for every edge e of the tree 4 arbitrary probabilities which sum up to 1, i.e. 3 free parameters, which may be different on different edges. Kimura's 2-parameter model is similar, but further restricted by $p_e(\{2\}) = p_e(M)$ for all edges, and finally, the Jukes-Cantor model requires in addition $p_e(\{1\}) = p_e(M)$ for all edges.

It is very interesting, that the models above were equipped with substitution mechanisms for transitions and transversions that fit perfectly the group theoretical description, although this was not the motivation for their invention.

The model, in which we work, slightly differs from Kimura's models, namely, we do not have a root r for an unknown common ancestor. This is in no way a serious loss, since, as we have already explained, it easily can be found by outgroup comparison. The root that we use, is, like in Section 3, *one arbitrary leaf R*, which represents an existing species. At every site of the sequence of R, we find a group element, and for normalization, in every leaf we multiply at the same site with the inverse of that group element. We refer to the sequences obtained as *normalized sequences*, note, that the normalized sequence of R contains identity elements only. From the normalized sequences we can read a leaf colouration at every bit; we count relative frequencies of leaf colourations and we treat these relative frequencies as if they were the f_{σ_1,σ_2} leaf colouration probabilities from the model of Section 3. Observe that the propagation of group elements along the tree is direction dependent unless $p_e(g) = p_e(g^{-1})$ for all e and g; and without this condition the normalization would not make sense. However, for $G = Z_2^m$, the condition holds automatically.

We had a set of species with corresponding segments of aligned DNA sequences. We selected an arbitrary species for R and we normalized the sequences from R, and obtained an f'_{σ_1,σ_2} relative frequency of the colouration (σ_1, σ_2) among the bits. Now we face the following problem: which tree T and probability distributions $p_e(g)$ over its edges yield a leaf colouration probability $f_{\sigma_1,\sigma_2} = f'_{\sigma_1,\sigma_2}$ for all (σ_1, σ_2)? Working with real data, we must be satisfied with the best approximation in a reasonable norm. Having the p_e's on the edges of the true tree allows for estimating a time scale, i.e. how far ago in time the evolutionary events in question did happen. The following theorem will give a solution for the problem; we formulate it for $G = Z_2^m$.

Let H denote the connecting Hadamard matrix in (8). Let $\mathbf{f}$ denote the vector of $f_{\sigma_1,\ldots,\sigma_m}$'s in (8). We adopt the convention of writing $[\mathbf{v}]_j$ for the j^{th} coordinate of the vector $\mathbf{v}$. Let K denote the Hadamard matrix, in which rows and columns are indexed with subsets of M, and the general h, g entry is

$$(-1)^{|h \cap g|};$$

let $\mathbf{p}_e$ denote the vector, for which $[\mathbf{p}_e]_h = p_e(h)$. For a positive vector $\mathbf{v}$, we denote by $\log \mathbf{v}$ the vector, for which $[\log \mathbf{v}]_i = \log[\mathbf{v}]_i$. We define an important set here, which is essential also for our results on invariants:

$$\mathcal{C}(T) = \left\{ (\sigma_1, ..., \sigma_m) : \ e \in E(T), \ h \subseteq M, \ \sigma_i = \begin{cases} L_e(T), & \text{if } i \in h, \\ \sigma_i = \emptyset, & \text{otherwise} \end{cases} \right\}. \tag{10}$$

We generalized with Hendy [SHSE] the spectral analysis/closest tree method as follows:

Theorem. *In the model of Section 3 for $G = Z_2^m$,*

$$[H^{-1} \log H\mathbf{f}]_{\sigma_1,\ldots,\sigma_m} =$$
$$\begin{cases} 0, & \text{if } (\sigma_1, ..., \sigma_m) \notin \mathcal{C}(T), \\ [K^{-1} \log K\mathbf{p}_e]_h, & \text{if } (\emptyset, ..., \emptyset) \neq (\sigma_1, ..., \sigma_m) \in \mathcal{C}(T) \\ & \qquad \text{defined by } e \text{ and } h \text{ in (10)}, \\ -\sum_{e \in E(T)} \sum_{\emptyset \neq h \in M} [K^{-1} \log K\mathbf{p}_e]_h, & \text{if } (\sigma_1, ..., \sigma_m) = (\emptyset, ..., \emptyset), \end{cases} \tag{11}$$

if all the logarithms are to be taken of positive numbers.

We note here, that the model of Section 3 does not imply the existence of the logarithms; however, for real data, there is no problem with them, due to the fact, that the probabilities $p_e(g)$ are sufficiently small for $g \neq (\emptyset, ..., \emptyset)$. Working with $\mathbf{f}$ arising from the model of Section 3, (11) and (10) tell the edges of the tree, and from (11) one can obtain $\mathbf{p}_e$ for all edges as well.

Working with empirical $\mathbf{f}'$, the closest tree method, which is a branch-and-bound algorithm, determines then the evolutionary tree and the $\mathbf{p}_e$'s over its edges, which yields $\mathbf{f}$, such that $H^{-1}\log H\mathbf{f}$ approximates $H^{-1}\log H\mathbf{f}'$ best in the Euclidean norm. The actual computation can be facilitated by writing H into a symmetric form achieving $H^{-1} = 4^{1-n}H$ and by an adaptation of the fast Fourier transform. The inverse pair (8)-(9) is a necessary tool in proving (11).

5. Invariants

There is a continuing interest in the theory of invariants of evolutionary trees. Roughly speaking, an invariant is a polynomial identity, which holds on one evolutionary tree no matter what the probabilities assigned to the edges are, and usually does not hold on other evolutionary trees. The great advantage of using invariants is that one may discriminate against some trees without (strong) assumptions regarding the probabilities. Invariants were introduced by Cavender and Felsenstein [CF], [C2], [C3] and Lake [L]; and recently Evans and Speed [EvS] gave an algebraic technique based on Fourier analysis to decide if a polynomial is invariant or not for Kimura's 3-parameter model.

Here we give explicitly a complete set of invariants for the mathematical model described in Sections 3-4 for $G = Z_2^m$. We still do not assume, that the tree is binary, but we assume, that no vertex has degree two. For a formal definition, let us be given a tree T and another tree T' on the same leaf set L and root R. Introduce the indeterminates $x_{\sigma_1,...,\sigma_m}$ for all $\sigma_i \subseteq L \setminus \{R\}$, $i = 1, 2, ..., m$. A multivariate polynomial $q(..., x_{\sigma_1,...,\sigma_m}, ...)$ is an *invariant* of the tree T, if q vanishes after the substitution of $f^T_{\sigma_1,...,\sigma_m}$'s into $x_{\sigma_1,...,\sigma_m}$'s, for any ξ_e independent random variables over the edges of T. We expect from an invariant, that it is non-zero on a typical $f^{T'}_{\sigma_1,...,\sigma_m}$; and hence searching for the true tree T', having the observed $f^{T'}_{\sigma_1,...,\sigma_m}$, we may reject a wrong candidate T, using its invariant(s).

A set of invariants of T is *complete*, if for any other tree T', at least one of the polynomials does not vanish on some $f^{T'}_{\sigma_1,...,\sigma_m}$. (Then, it comes for free, that it discriminates against almost all probability distributions.) For $m = 1$, our complete set of invariants was already found by Hendy [H], although it is not explicit there. These invariants are very similar to (11), but note, that (11) is not a polynomial identity. Define the polynomials

$$R_{X_1,...,X_m} = \sum_{\sigma_1,...,\sigma_m} (-1)^{\sum_{i=1}^m |\sigma_i \cap X_i|} x_{\sigma_1,...,\sigma_m}$$

for $X_i \subseteq L$, $|X_i|$ even, $i = 1, 2, ..., m$. Now for an arbitrary given $(\rho_1, ..., \rho_m)$ ($\rho_i \subseteq L \setminus \{R\}$, $i = 1, 2, ..., m$), define the polynomial $\delta_{\rho_1,...,\rho_m}$ of all the variables $x_{\sigma_1,...,\sigma_m}$:

$$\delta_{\rho_1,...,\rho_m} = \prod_{\substack{(X_1,...,X_m): \\ \sum_{i=1}^m |X_i \cap \rho_i| \equiv 0 \ mod \ 2}} R_{X_1,...,X_m} - \prod_{\substack{(X_1,...,X_m): \\ \sum_{i=1}^m |X_i \cap \rho_i| \equiv 1 \ mod \ 2}} R_{X_1,...,X_m}.$$

Theorem. *The polynomials* $\{\delta_{\rho_1,...,\rho_m} : (\rho_1, ..., \rho_m) \notin \mathcal{C}(T)\}$ *make a complete set of invariants of* T.

It is worth making the following comment here. Evans and Speed [EvS] made the following conjecture: "the number of algebraically independent invariants and the number of free parameters among the $p_e(g)$'s obtained by an informal parameter count add up to the number of variables $x_{\sigma_1,...,\sigma_m}$". Their first problem seems to have been to set candidates for these independent invariants. Assume that for $g \neq \emptyset$, $p_e(g)$ is a variable and $p_e(\emptyset) = 1 - \sum_{g \neq \emptyset} p_e(g)$, then the number of variables $f_{\sigma_1,...,\sigma_m}$ is $2^{m(n-1)}$, the number of free parameters is $|E(T)|(2^m - 1)$, the number of invariants given in the theorem is $2^{m(n-1)} - |\mathcal{C}(T)| = 2^{m(n-1)} - |E(T)|(2^m - 1) - 1$; and actually, we have one more invariant, $\sum f_{\sigma_1,...,\sigma_m} = 1$. The numerology works, but a positive result here would seem to involve algebraic geometry.

If it comes to application of these invariants, then values of polynomial functions must be computed instead of the polynomials, since computer algebra in many variables is rather prohibitive.

Problem. *Generalize the above set of invariants to the case of arbitarary finite Abelian group.*

6. Conclusion

The spectral analysis method has the advantage of using all the genetic information from the sequences, a property, which is not shared by most other reconstruction techniques. As it was pointed out in [H], [PHS1], [PHS2], it satisfies the Popperian program of falsifiability. Namely, the probabilities $p_e(h)$ resulting from (11) might be negative numbers in the closest tree. That this can happen for artificial data but not for real data is a circumstancial evidence for the truth of Cavender's model and Kimura's 3-parameter model. There is an additional Popperian test for Kimura's 3-parameter model, namely, that in (11), for $(\sigma_1, \sigma_2) \notin \mathcal{C}(T)$, $\sigma_1 \neq \sigma_2$, $[H^{-1} \log H\mathbf{f}]_{\sigma_1,\sigma_2} = 0$; and this test does not even assume any knowledge on the closest tree.

Compared with spectral analysis, the parsimony principle is a rather rough exploratory method. However, with small binary trees and uniform small probabilities $p_e(g)$ for any change ($g \neq$identity), $p_e(g)^2 << p_e(g)$, changing twice for a nucleotide is highly unlikely, and the parsimony principle turns into an approximation of Kimura's model. The parsimony principle and the closest tree method are both minimum principles, although with different objective functions.

The second author proposes the development of randomized algorithms for tree reconstruction. In view of the successes of randomized algorithms in situations where deterministic algorithms fail, this approach could be promising, although nothing is done yet.

REFERENCES

[C1] J. A. Cavender, Taxonomy with confidence, *Math. Biosci.* **40**(1978), 271–280.

[C2] J. A. Cavender, Mechanized derivations of linear invariants, *Mol. Biol. and Evol.* **6**(1989), 301–316.

[C3] J. A. Cavender, Necessary conditions for the method of inferring phylogeny by linear invariants, *Math. Biosci.* **103**(1991), 69–75.

[CF] J. A. Cavender and J. Felsenstein, Invariants of phylogenies in a simple case with discrete states, *J. Class.* **4**(1987), 57–71.

[CHPSW] M. Carter, M. D. Hendy, D. Penny, L. A. Székely, N. C. Wormald, On the distribution of length of evolutionary trees, *SIAM J. Discrete Math.* **3**(1990), 38–47.

[CSE] L. L. Cavalli-Sforza, A. W. F. Edwards, Phylogenetic analysis: models and estimation procedures, *Evolution* **21**(1967), 550–570.

[DJPSY] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriu, P. Seymour, M. Yannakakis, The complexity of multiway cuts, Extended abstract, 1983.

[E] P. L. Erdős, A new bijection on rooted forests, in: *Proceedings of the* 4^{th} *French Combinatorial Conference, Marseille, 1990*, to appear.

[ES1] P. L. Erdős, L. A. Székely, Applications of antilexicographic order I: An enumerative theory of trees, *Adv. Appl. Math.* **10**(1989), 488–496.

[ES2] P. L. Erdős, L. A. Székely, Counting bichromatic evolutionary trees, to appear in *Discrete Appl. Math.*

[ES3] P. L. Erdős, L. A. Székely, Evolutionary trees: An integer multicommodity max-flow–min-cut theorem, to appear in *Adv. Appl. Math.*

[EvS] S. N. Evans, T. P. Speed, Invariants of some probability models used in phylogenetic inference, manuscript.

[F] J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.* **27**(1978), 401–410.

[FG] L. R. Foulds, R. L. Graham, The Steiner problem in phylogeny is NP-complete, *Adv. Appl. Math.* **3**(1982), 43–49.

[H] M. D. Hendy, A combinatorial description of the closest tree algorithm for finding evolutionary trees, *Discrete Math.* **96**(1991), 51–58.

[HHP] I. M. Henderson, M. D. Hendy, D. Penny, Influenza viruses, comets and the science of evolutionary trees, *J. Theor. Biol.* **140**(1989), 289–303.

[HP1] M. D. Hendy, D. Penny, A framework for the quantitative study of evolutionary trees, *Systematic Zoology* **38**(4) (1989), 297–309.

[HP2] M. D. Hendy, D. Penny, Spectral analysis of phylogenetic data, preprint, University of Bielefeld, ZiF-Nr. 91/23.

[J] N. Jacobson, *Basic Algebra II,* W. H. Freeman and Co. San Francisco, 1980.

[JC] T. H. Jukes, C. Cantor, Evolution in protein molecules, in: *Mammalian Protein Metabolism* (H. N. Munro, ed.), 21–132, New York, Academic Press, 1969.

[K1] M. Kimura, A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences, *J. Mol. Evol.* **16**(1980), 111–120

[K2] M. Kimura, Estimation of evolutionary sequences between homologous nucleotide sequences, *Proc. Natl. Acad. Sci. USA* **78**(1981), 454–458.

[K3] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, 1983.

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

[N] J. Neyman, Molecular studies of evolution: A source of novel statistical problems, in: *Statistical Decision Theory and Related Topics*, (S. S. Gupta and J. Yackel, eds.) 1–27, New York, Academic Press, 1971.

[PHS1] D. Penny, M. D. Hendy, M. A. Steel, Progress with methods for constructing evolutionary trees, *Trends in Ecology & Evolution* **7**(1992)(3), 73–79.

[PHS2] D. Penny, M. D. Hendy, M. A. Steel, Testing the theory of descent, in: *Phylogenetic Analysis of DNA Sequences*, eds. M. M. Miyamoto, J. Cracraft, Oxford University Press, New York–London, 1991, 155–183.

[PHZH] D. Penny, M. D. Hendy, E. A. Zimmer, R. K. Hamby, Trees from sequences: panacea or Pandora's box? *Aust. Syst. Bot.* **3**(1990), 21–38

[S1] M. A. Steel, Distributions on bicoloured binary trees arising from the principle of parsimony, to appear in *Discrete Appl. Math.*

[S2] M. A. Steel, Decompositions of leaf-coloured binary trees, to appear in *Adv. Appl. Math.*

[SHP] M. A. Steel, M. D. Hendy, D. Penny, Significance of the length of the shortest tree, *J. Classification* **9**(1992), 71–90.

[SHSE] M. A. Steel, M. D. Hendy, L. A. Székely, P. L. Erdős, Spectral analysis and a closest tree method for genetic sequences, subbmitted to *Appl. Math. Letters.*

[SESP] L. A. Székely, P. L. Erdős, M. A. Steel, D. Penny, A Fourier inversion formula for evolutionary trees, submitted to *Appl. Math. Letters.*

[SSE] L. A. Székely, M. A. Steel, P. L. Erdős, Fourier calculus on finite sets and evolutionary trees, submitted to *Discrete Math.*

L. A. Székely, Eötvös University, Budapest, Hungary

and Institut für diskrete Mathematik, Bonn

P. L. Erdős, Hungarian Academy of Sciences, Budapest, Hungary

M. A. Steel, University of Canterbury, Christchurch, New Zealand