1992

L.A. Székely, P.L. Erdös, M.A. Steel

The combinatorics of reconstructing evolutionary trees

Department of Operations Research, Statistics, and System Theory Report BS-R9233 December

CWI is het Centrum voor Wiskunde en Informatica van de Stichting Mathematisch Centrum
CWI is the Centre for Mathematics and Computer Science of the Mathematical Centre Foundation

CWI is the research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a non-profit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands organization for scientific research (NWO).

Copyright © Stichting Mathematisch Centrum, Amsterdam

The Combinatorics of Reconstructing Evolutionary Trees

L. A. Székely, P. L. Erdős, M. A. Steel
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Abstract

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

1991 Mathematics Subject Classification: 92D15, 92B10, 05C05, 42A61
Key words and Phrases: Evolutionary trees, maximum parsimony, Fourier transform, Kimura's 3-parameter model;
Note: Research of the first author was supported by the A. v. Humboldt-Stiftung while he was at the Rheinische Friedrich-Wilhelms Universität, Bonn, and the U. S. Office of Naval Research under the contract N-0014-91-J-1385. Research of the second author was done while visiting CWI.

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.

Report BS-R9233
ISSN 0924-0659
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

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 of nucleotides. Biologists also would like to add a root r to denote a common ancestor and the direction of the evolution. This root r may subdivide an edge of the true tree or may be attached to a vertex of the true tree. However, if you have a procedure to solve the problem above without r, 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 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. A preliminary version of the present paper appeared in the conference proceedings [SES].

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. 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, Section 6 concludes with remarks on algorithmic and philosophical aspects.

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

We need to recall some facts on characters and the Fourier transform, which can be found in [J] or in [EvS]. We use additive notation in Abelian groups.

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 sums of characters.

Let us be given a tree T with leaf set L and one arbitrary leaf R, called a *root*. We assume that no vertex has degree two. Assume 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 distributions $p_e(g) := Prob(\xi_e = g)$, such that $\sum_{g \in G} p_e(g) = 1$. We call the set of p_e distributions ($e \in E(T)$) a *transition mechanism* and denote it by p.

Take G^{n-1} = the set of leaf colourations $\sigma : L \setminus \{R\} \longrightarrow G$ endowed with pointwise operation; we denote the value of σ at l by σ_l . 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 sum of group elements along the unique Rl path. Let f_σ denote the probability that we obtain the leaf colouration $\sigma : L \setminus \{R\} \longrightarrow G$ in this way. In case we want to emphasize the dependence from the tree T and the transition mechanism p, we will write $f_\sigma(T, p)$.

Let $\chi = (\chi_l \in \hat{G} : l \in L \setminus \{R\})$ be an ordered $(n-1)$-tuple of characters. Then $\chi \in \hat{G}^{n-1}$, and χ acts on G^{n-1} according to Lemma (iii). For $e \in E(T)$, set

$$L_e = \{l \in L : e \text{ separates } l \text{ from } R \text{ in } T\}.$$

For $e \in E(T)$ and $\chi \in \hat{G}^{n-1}$, set $\chi_e = \sum_{l \in L_e} \chi_l$, so $\chi_e \in \hat{G}$. For $h \in \hat{G}$, $e \in E(T)$ define $l_e(h) = \sum_{g \in G} h(g)p_e(g)$, and

$$r_\chi = \prod_{e \in E(T)} l_e(\chi_e). \tag{5}$$

In [SSE] we obtained the following Fourier inverse pair:

Theorem. *With* $\chi(\sigma) = \prod_{l \in L \setminus \{R\}} \chi_l(\sigma_l)$,

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

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

In [SSE] we observed that (6) and (7) are equivalent by Lemma (ii) for *any* f and r; and it is not difficult to prove (6) for *our* f_σ and r_χ, based on the factorization (5).

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 $G = Z_2$, 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 $G = Z_2 \times Z_2$, i.e. for sequences over the 4-letter alphabet A, G, C, T. We explain the $G = Z_2 \times Z_2$ case in details, the explanation also applies, mutatis mutandis, to $G = Z_2$. 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 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. We follow the group theoretical setting of the models from Evans and Speed [EvS]. Identify the elements of $Z_2 \times Z_2$ with the four nucleotides, such that A= 0. Take the true tree with a common ancestor r, assume that an element of $Z_2 \times Z_2$ is assigned under a certain (unknown) distribution to r. The random group element at r is regarded as the original nucleotide value there. To every edge of the tree a random element of $Z_2 \times Z_2$ is assigned independently, the distribution may vary from edge to edge. The random variable at an edge describes the nucleotide change on that edge. In terms of biology, adding A=0 on an edge causes no change in the nucleotide, adding G causes *transition*, and adding C or T causes one of the two possible types of *transversions*. To every leaf l the sum 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 $p_e(\mathrm{G}) = p_e(\mathrm{T})$ for all edges, and finally, the Jukes-Cantor model requires in addition $p_e(\mathrm{C}) = p_e(\mathrm{T})$ for all edges.

It is surprising enough, 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 it is easy to recover it 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 standardization, in every leaf we multiply at the same site with the inverse of that group element. We refer to the sequences obtained as *standardized sequences*, note, that the standardized sequence of R contains 0's only. From the standardized sequences we can read a leaf colouration at every bit; we count relative frequencies of leaf colourations and we treat these relative frequencies f'_σ as if they were the f_σ leaf colouration probabilities from the model of Section 2. 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 standardization would not make sense. However, for $G = Z_2^m$, the condition holds automatically. Standardization sets no restriction on the distribution at r, since we rather work with nucleotide changes than use the nucleotide values. Despite the small difference, our method will allow for reconstruction of the true tree that evolved according to Kimura's model, with the loss of r and with the possible loss of the vertex adjacent to r, if it has degree 3.

Now we face the following problem: which tree T and probability distributions $p_e(g)$ over its edges yield a leaf colouration probability $f_\sigma = f'_\sigma$ for all σ? One easily sees, that with $p_e(g) = 1/|G|$ for all $e \in E(T)$ and $g \in G$, all leaf colourations are equally likely, independently of the shape of the tree. Hence, there is no way to reconstruct the tree. However, the following theorem shows, that reconstruction *is* possible, if $p_e(0)$ (i.e. the probability of no change) is sufficiently close to 1 on all edges. Fortunately, this is the case with evolution.

Let H denote the connecting matrix in (6), i.e. the rows correspond to elements of $\hat{G}^{n-1}$, the columns correspond to elements of G^{n-1}, and the (χ, σ) entry is $\chi(\sigma)$. Let $\mathbf{f}$ denote the vector of f_σ's in (6). We adopt the convention of writing $[\mathbf{v}]_j$ for the j^{th} coordinate of the vector $\mathbf{v}$. Let K denote the matrix, in which rows correspond to elements of $\hat{G}$ and columns correspond to the elements of G, and the (h, g) entry is $h(g)$. 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. For $e \in E(T)$, $0 \neq g \in G$, define $\rho^{e,g} \in G^{n-1}$ in the

following way: $\rho_l^{e,g} = 0$ for $l \notin L_e$, $l \neq R$, and $\rho_l^{e,g} = g$ for $l \in L_e$. Define

$$\mathcal{C}(T) = \{\rho^{e,g} : e \in E(T), 0 \neq g \in G\}. \tag{8}$$

In [SHSE]and [SSE] we generalized the spectral analysis/closest tree method as follows:

Theorem. *If $p_e(0)$ is sufficiently close to 1, then*

$$[H^{-1} \log H\mathbf{f}]_\rho = \begin{cases} 0, & \text{if } 0 \neq \rho \notin \mathcal{C}(T), \\ [K^{-1} \log K\mathbf{p}_e]_h, & \text{if } \rho = \rho^{e,h} \in \mathcal{C}(T), \\ \sum_{e \in E(T)} [K^{-1} \log K\mathbf{p}_e]_0, & \text{if } \rho = 0. \end{cases} \tag{9}$$

We use complex logarithm in a neighbourhood of 1 such that $\log 1 = 0$. For real data, due to the fact that $p_e(0)$ is sufficiently small, we hit this neighbourhood. Working with $\mathbf{f}$ arising from the model of Section 3, (8) and (9) tell the edges of the tree, and from (9) one can obtain $\mathbf{p}_e$ for all edges as well.

Working with empirical $\mathbf{f}'$, 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 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 closest tree method for $Z_2 \times Z_2$, i.e. for four character state sequence, was already succesfully applied to real data [HPS].

The proof of (9) in [SSE] is purely combinatorial, the inverse pair (6)-(7) and the factorization (5) is a necessary tool in it.

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 transition mechanism is, 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.

Let us be given a tree T and another tree T' on the same leaf set L and root R. Consider the indeterminates x_σ for $\sigma \in G^{n-1}$ again. A multivariate function $q_T(...,x_\sigma,...)$ is an *invariant* of the tree T, if q vanishes after the substitution of $f_\sigma(T,p)$'s into x_σ's, for any transition mechanism p of T. We expect that an invariant is non-zero for a typical substitution of $f_\sigma(T',p')$'s into the x_σ's; and hence searching for the tree T' and its transition mechanism p' that resulted in the observed f_σ, we may reject a wrong candidate T, using its invariant(s).

Consider

$$Split(T) = \left\{ L_e(T) : e \in E(T) \right\}$$

and observe that every element of $Split(T)$ is represented by a *unique* edge e, since T has no vertex of degree two. Call an edge $e \in E(T)$ *passive* for (T,p), if $p_e(0) = 1$. Consider the set of ordered pairs (trees, transition mechanisms) on the same fixed leaf set L and root R; and define a relation $\sim$ by $(T,p) \sim (T',p')$ iff a (T'',p'') can be reached from both by contracting passive edges. It is easy to see that $\sim$ is an equivalence relation.

Define the polynomial $R_\chi = \sum_{\sigma \in G^{n-1}} \chi(\sigma) x_\sigma$ for $\chi \in \hat{G}^{n-1}$. For $\rho \in G^{n-1}$, define the tree independent $C^n \to C$ functions

$$\delta_\rho = \prod_{\chi \in \hat{G}^{n-1}} R_\chi^{\overline{\chi(\rho)}} - 1$$

in a neighborhood of $x_0 = 1$, $x_\sigma = 0$. For $0 \neq \rho \notin C(T)$, we term the δ_ρ's as the *canonical invariants* of the tree T. (It is easy to derive from (9) that they are invariants indeed.)

For the main results of this Section we put $p_e(0)$ into the first coordinate in $\mathbf{p}_e$.

Theorem. *Assume that for the transition mechanisms p and p', for any edge e the vectors $\mathbf{p}_e$ and $\mathbf{p}_e{}'$ are sufficiently close to $(1,0,...,0)^T$.*
(i) If $f_\sigma(T,p)$ satisfies the canonical invariants of T', then the elements of $Split(T) \setminus Split(T')$ are represented by passive edges in T.
(ii) If $f_\sigma(T,p)$ satisfies the canonical invariants of T' and $f_\sigma(T',p')$ satisfies the canonical invariants of T, then $(T,p) \sim (T',p')$.
(iii) If a leaf colouration probability distribution f_σ comes from both (T,p) and (T',p'), then $(T,p) \sim (T',p')$.
(iv) The canonical invariants of the tree T are algebraically independent.

In the rest of the Section we restrict ourselves to $G = Z_2^m$. For an arbitrary given $\rho \in (Z_2^m)^{n-1}$, we define the polynomial δ'_ρ of all x_σ's:

$$\delta'_\rho = \prod_{\substack{\chi \in (\widehat{Z_2^m})^{n-1}: \\ \chi(\rho)=1}} R_\chi - \prod_{\substack{\chi \in (\widehat{Z_2^m})^{n-1}: \\ \chi(\rho)=-1}} R_\chi.$$

Clearly, we obtained polynomial invariants, of which most of the theorem can be easily told, with the annoying exception of their algebraic independence. In fact, we conjecture that the polynomials δ'_ρ altogether with the polynomial $R_0 - 1 = \sum_\sigma x_\sigma - 1$ are algebraically independent.

It is worth making the following comment here. Evans and Speed [EvS] conjecture that "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_σ". Their first problem seems to have been to set candidates for these independent invariants. We have the suggestion above. Assume that for $g \neq 0$, $p_e(g)$ is a variable and $p_e(0) = 1 - \sum_{g \neq 0} p_e(g)$; then the number of free parameters is $|E(T)|(2^m - 1)$, the number of variables x_σ is $2^{m(n-1)}$, the number of canonical invariants δ'_ρ is $2^{m(n-1)} - |C(T)| - 1 = 2^{m(n-1)} - |E(T)|(2^m - 1) - 1$; and actually, we have one more invariant, $R_0 - 1 = \sum_\sigma x_\sigma - 1$. The numerology works, but a positive result here would seem to involve algebraic geometry. We gave some support for the conjecture.

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 (9) 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 (9), for $\sigma \notin C(T)$, $[H^{-1} \log H\mathbf{f}]_\sigma = 0$; 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, if the Jukes-Cantor model or Cavender's model applies to a small binary

tree such that there are small equal changing probabilities $p_e(g) = p$ ($g \neq$identity) on all edges, then we have $p_e(g)^2 << p_e(g)$, and changing twice for a nucleotide is highly unlikely; in such circumstances the parsimony principle is expected to yield the true tree. The parsimony principle and the closest tree method are both minimum principles, although with different objective functions.

It is appropriate to comment here on the computational complexity issue. Clearly, working with $4^{n-1} \times 4^{n-1}$ matrices in order to reconstruct a tree on n leaves is not computationally feasible for large values of n. There are, however, polynomial time algorithms to reconstruct the evolutionary tree, if a consistent method can determine the true tree for any 4-subset of species. These methods may not be reliable on real data and do not provide for the transition mechanism.

We suggest here a polynomial time algorithm based on (9), that we expect to be reliable and computationally feasible at the same time. We do not estimate the running time, since it may depend on the implementation. The algorithm is based on three observations:

(i) in (9), we only want to compute the coordinates corresponding to a $\rho^{e,h}$, where e defines a split of the tree (but we do not know in advance, which coordinates they are),

(ii) the number of different leaf colourations that occur in the data seems to be $O(n^2)$ by experience, but in no way can exceed the length of the genetic sequences considered (much less than 4^{n-1}),

(iii) a second order approximation formula for (9) ([SSE]): for $\mathbf{x}$ with $x_\sigma \geq 0$ and $\sum x_\sigma = 1$, for $\sigma \neq 0$,

$$[H^{-1} \log H\mathbf{x}]_\sigma \approx \frac{x_\sigma}{x_0} - \frac{1}{2} \sum_{\substack{(\sigma_1,\sigma_2): \\ \sigma_1+\sigma_2=\sigma \\ \sigma_1,\sigma_2 \neq 0}} \frac{x_{\sigma_1} x_{\sigma_2}}{x_0^2}.$$

Now the algorithm would apply (iii) for certain subsets L' of L to decide if a certain bipartition of L' is a split of the true tree on L'. Any split of L' is the trace of a split of L by the course of the evolution. Hence, having a split of L', one can blow it up into a split of L, by repeatedly adding a new vertex to some side of the split. Again, (iii) tests to which side the new vertex is to go. The algorithm may start with the split $a|R$ of $L' = \{a, R\}$. Having obtained a split of L, we may apply the same algorithm recursively to the vertex sets on the two sides of the split, until we recover all splits of L and use (iii) to recover the transition mechanism on the edges of the true tree.

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 in *Discrete Math.*

[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, in press, *Annals of Statistics*

[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.

[HPS] M. D. Hendy, D. Penny, M. A. Steel, Discrete Fourier analysis for evolutionary trees, submitted to *Proc. Natl. Acad. Sci. USA.*

[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, to appear in *Appl. Math. Letters.*

[SES] L. A. Székely, P. L. Erdős, M. A. Steel, The combinatorics of evolutionary trees—a survey, in: *Actes du Seminaire, Séminaire Lotharingien de Combinatoire, 28-ième session, 15–18 mars, 1992*, D. Foata, éd., Publication de l'Institute de Recherche Mathématique Avancée.

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

[SSE] L. A. Székely, M. A. Steel, P. L. Erdős, Fourier calculus on evolutionary trees, to appear in *Adv. Appl. Math.*

L. A. Székely, Department of Computer Science, Eötvös University, Múzeum krt. 6-8, 1088 Budapest, Hungary and University of New Mexico, Albuquerque, NM 87131-1141, U.S.A.

P. L. Erdős, Hungarian Academy of Sciences, Reáltanoda u. 13-15, 1053 Budapest, Hungary and Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

M. A. Steel, Department of Mathematics, University of Canterbury, Christchurch 1, New Zealand