FOURIER CALCULUS ON EVOLUTIONARY TREES

by

L.A. Székely,* M.A. Steel, P.L. Erdös

No. 70 August, 1992.

Abstract

We describe a Fourier analysis approach to the reconstruction theory of evolutionary trees that is based on Kimura's model of molecular evolution.

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

Fourier Calculus on Evolutionary Trees

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

ABSTRACT: We describe a Fourier analysis approach to the reconstruction theory of evolutionary trees that is based on Kimura's model of molecular evolution.

1. Introduction

The purpose of the present paper is to develop in full generality the mathematical tools that are being used in the spectral analysis/closest tree method [H], [HP1], [HP2], [SESP], [SHSE], [HPS] for the reconstruction of evolutionary trees in Cavender's model [C1] and in Kimura's 3-parameter model [K1], [K2], [K3]. All sections of this paper but the very last can be read with zero knowledge from biology. The last section explains the biological significance of the results from previous sections. An important tool of our work is the Fourier calculus over finite Abelian groups; we acknowledge the influence of Evans and Speed [ES]. We have already announced part of the results of the present paper without proofs in [SES]. The following lemma summarizes the basic facts that we need on characters and Fourier transform. We use the additive notation in Abelian groups.

Lemma 1. *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 $g \in G$

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

* 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

(iii) *The characters of a direct product of finite Abelian groups are exactly the products of characters.*

Proof. See [Kö].∎

Assume $A = (a_{ij})$ is a $p \times q$ matrix with integer entries. Let us be given a finite Abelian group G and the elements of G^q written in a vector form $\mathbf{x} = (x_1, \ldots, x_q)^T$, where $x_j \in G$. Define the vector $\mathbf{y} \in G^p$ by $\mathbf{y} = (y_1, \ldots, y_p)^T$, such that

$$y_i = \sum_{j=1}^{q} a_{ij} x_j.$$

(We want to abbreviate this fact to $A\mathbf{x} = \mathbf{y}$ and do not abuse this formalism.) Let us be given $p_j : G \to C$ functions $(j = 1, ..., q)$. Define for $\mathbf{x} = (x_1, \ldots, x_q)^T \in G^q$,

$$F(\mathbf{x}) = \prod_{j=1}^{q} p_j(x_j).$$

For $\mathbf{y} = (y_1, \ldots, y_p)^T \in G^p$, let

$$f(\mathbf{y}) = \sum_{\substack{\mathbf{x} \in G^q \\ A\mathbf{x} = \mathbf{y}}} F(\mathbf{x}).$$

Theorem 2. *If* $\chi = (\chi_1, \ldots, \chi_p)^T \in \hat{G}^p$, *then*

$$\hat{f}(\chi) = \prod_{j=1}^{q} \sum_{x \in G} p_j(x) \Big(\sum_{i=1}^{p} a_{ij} \chi_i\Big)(x).$$

Proof. By definition,

$$\hat{f}(\chi) = \sum_{\mathbf{y} \in G^p} \chi(\mathbf{y}) f(\mathbf{y}) = \sum_{\mathbf{y} \in G^p} \chi(\mathbf{y}) \sum_{\substack{\mathbf{x} \in G^q : \\ A\mathbf{x} = \mathbf{y}}} F(\mathbf{x}) = \sum_{\mathbf{x} \in G^q} F(\mathbf{x}) \chi(A\mathbf{x}).$$

Now we have

$$\chi(A\mathbf{x}) = \prod_{i=1}^{p} \chi_i((A\mathbf{x})_i) = \prod_{i=1}^{p} \chi_i\Big(\sum_{j=1}^{q} a_{ij} x_j\Big) = \prod_{j=1}^{q} \prod_{i=1}^{p} \chi_i(a_{ij} x_j).$$

Hence,

$$\hat{f}(\chi) = \sum_{\mathbf{x} \in G^q} \prod_{j=1}^{q} p_j(x_j) \prod_{i=1}^{p} \chi_i(a_{ij} x_j) = \prod_{j=1}^{q} \sum_{x \in G} p_j(x_j) \prod_{i=1}^{p} \chi_i(a_{ij} x_j),$$

as claimed. ∎

Note that for $A = [1, 1]$, $\mathbf{x} = (f, g)^T$, Theorem 2 gives back a special instance of the classical result for the Fourier transform of the convolution, $\widehat{f * g} = \hat{f} \cdot \hat{g}$.

2. Our model and its basic identities

First we describe the mathematical model, which we work with. 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 1(*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

$$\chi_e = \sum_{l \in L_e} \chi_l, \tag{1}$$

so $\chi_e \in \hat{G}$. For $h \in \hat{G}$, $e \in E(T)$ define

$$l_e(h) = \sum_{g \in G} h(g) p_e(g), \quad \text{and} \tag{2}$$

$$r_\chi = \prod_{e \in E(T)} l_e(\chi_e). \tag{3}$$

We have the following Fourier inverse pair:

Theorem 3. *With* $\chi(\sigma) = \prod_{l \in L \setminus \{R\}} \chi_l(\sigma_l)$,

$$r_\chi = \sum_{\sigma \in G^{n-1}} \chi(\sigma) f_\sigma \qquad \textit{and} \tag{4}$$

$$f_\sigma = \frac{1}{|G|^{n-1}} \sum_{\chi \in \hat{G}^{n-1}} \overline{\chi(\sigma)} r_\chi. \tag{5}$$

Proof. Observe that (4) and (5) are equivalent by Lemma 1(*ii*) for *any* $f : G^{n-1} \longrightarrow C$ and $r : \hat{G}^{n-1} \longrightarrow C$. We decided not to use the usual hat notation for this pair since their significance and frequent occurence in this paper. To prove (4) with *our* f_σ and r_χ, apply Theorem 2 in the following setting: $p = n-1$, $q = |E(T)|$, $A = (a_{ie})$ with

$$a_{ie} = \begin{cases} 1 & \text{if edge } e \text{ lies on the } Ri \text{ path} \\ 0 & \text{otherwise.} \end{cases}$$

Take $\Xi = (\xi_e : e \in E(T))$ the vector of random group elements selected independently on the edges, $p_e(x) = Prob(\xi_e = x)$, $\Upsilon =$ the vector of the resulting random leaf colouration. Observe that the independence implies $F(\mathbf{x}) = Prob(\Xi = \mathbf{x})$, and $f(\mathbf{y}) = Prob(\Upsilon = \mathbf{y})$. ∎

For later use we define the polynomials $R_\chi = \sum_{\sigma \in G^{n-1}} \chi(\sigma) x_\sigma$, with independent variables x_σ. Observe that while R_χ is tree independent, $r_\chi = R_\chi|_{x_\sigma = f_\sigma}$ is tree dependent.

Theorem 4. *For the transition mechanisms* $p^{(i)}$, p^* *on the tree* T *and* $\sigma \in G^{n-1}$ *we have*

$$\sum_{\substack{(\sigma_1, \sigma_2, \ldots, \sigma_k): \\ \sigma_1 + \sigma_2 + \ldots + \sigma_k = \sigma \\ \sigma_i \in G^{n-1}}} \prod_{i=1}^{k} f_{\sigma_i}(T, p^{(i)}) = f_\sigma(T, p^*),$$

where for $g \in G$

$$p_e^*(g) = \sum_{\substack{(g_1, g_2, \ldots, g_k): \\ g_1 + g_2 + \ldots + g_k = g \\ g_i \in G}} \prod_{i=1}^{k} p_e^{(i)}(g_i).$$

Proof. Define for $\sigma \in G^{n-1}$

$$f(\sigma) = \sum_{\substack{(\sigma_1, \sigma_2, \ldots, \sigma_k): \\ \sigma_1 + \sigma_2 + \ldots + \sigma_k = \sigma \\ \sigma_i \in G^{n-1}}} \prod_{i=1}^{k} f_{\sigma_i}(T, p^{(i)})$$

and $f_i(\sigma) = f_\sigma(T, p^{(i)})$. We are going to prove $f(\sigma) = f_\sigma(T, p^*)$. Applying Theorem 2 to the group G^{n-1} in the setting $p = k$, $q = 1$, $A = (1, 1, \ldots, 1)$, $p_i(\sigma) = f_i(\sigma)$ yields

$$\hat{f}(\chi) = \prod_{i=1}^{k} \hat{f}_i(\chi);$$

and by Theorem 3 and (3)

$$\hat{f}_i(\chi) = \prod_{e \in E(T)} \sum_{g \in G} \chi_e(g) p_e^{(i)}(g).$$

Therefore,

$$\hat{f}(\chi) = \prod_{e \in E(T)} \sum_{g \in G} \chi_e(g) p_e^*(g).$$

Finally, by Theorem 3 .

$$\frac{1}{|G|^{n-1}} \sum_{\chi \in G^{n-1}} \overline{\chi(\sigma)} \hat{f}(\sigma) = f_\sigma(T, p^*),$$

and by Lemma $1(ii)$

$$f(\sigma) = \frac{1}{|G|^{n-1}} \sum_{\chi \in G^{n-1}} \overline{\chi(\sigma)} \hat{f}(\sigma);$$

yielding the wanted $f(\sigma) = f_\sigma(T, p^*)$.∎

We note that a special case of Theorem 4 occured in the Ph. D. Thesis of the second author, [S]. An algebra oriented reader may be interested in the fact, that Theorem 4 boils down to the commutative law in the group algebra $C[G^{n-1}]$.

3. Main identities

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 $\mathcal{C}(T) = \{\rho^{e,g} : e \in E(T), 0 \neq g \in G\}$. For the following theorem (and later on) we assume, that for every $e \in E(T)$, $p_e(0)$ is sufficiently close to 1, and hence r_χ is also sufficiently close to 1; and therefore logarithm (such that $\log 1 = 0$) can be given a satisfactory definition. Having the logarithm, complex exponentiation a^b will be $exp(b \log a)$, as usual.

Theorem 5. *For* $0_{G^{n-1}} \neq \rho \in G^{n-1}$, $\rho \notin \mathcal{C}(T)$,

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = 1;$$

for $\rho = \rho^{e,g} \in \mathcal{C}(T)$

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = \prod_{h \in \hat{G}} l_e(h)^{h(g)|G|^{n-2}};$$

and for $\rho = 0_{G^{n-1}}$

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = \prod_{e \in E(T)} \prod_{h \in \hat{G}} l_e(h)^{|G|^{n-2}}.$$

The identities remain valid with all exponents conjugated.

Proof. By (3) we have

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = \prod_{e \in E(T)} \prod_{h \in \hat{G}} l_e(h)^{\sum\{\chi(\rho) : \chi_e = h\}};$$

(1)-(2) altogether with $\chi(\rho) = \prod_{l \in L \setminus \{R\}} \chi_l(\rho_l)$ imply

$$\sum\{\chi(\rho) : \chi_e = h\} = \sum_{\substack{\chi_l \in \hat{G}: \\ l \in L \setminus \{R\}}} \left\{ \prod_{l \in L \setminus \{R\}} \chi_l(\rho_l) \ : \ \sum_{l \in L_e} \chi_l = h \right\}. \tag{6}$$

Now it is obvious that for $\rho = 0_{G^{n-1}}$

$$\sum_{\substack{\chi_l \in \hat{G}: \\ l \in L \setminus \{R\}}} \left\{ 1 : \sum_{l \in L_e} \chi_l = h \right\} = |G|^{n-2},$$

since having fixed an arbitrary $j \in L_e$, we have $|G|$ choices for χ_l for any $l \in L \setminus \{R, j\}$, and finally a unique choice for χ_j. Similarly, for $\rho = \rho^{e,g} \in \mathcal{C}(T)$

$$\sum_{\substack{\chi_l \in \hat{G}: \\ l \in L \setminus \{R\}}} \left\{ \prod_{l \in L \setminus \{R\}} \chi_l(\rho_l) : \sum_{l \in L_e} \chi_l = h \right\} = h(g)|G|^{n-2},$$

since for any $\chi = (\chi_l : l \in L \setminus \{R\})$, $\chi(\rho^{e,g}) = h(g)$, and having fixed an arbitrary $j \in L_e$, we have $|G|$ choices for χ_l for any $l \in L \setminus \{R, j\}$, and finally a unique choice for χ_j, like above.

The nontrivial part of the proof is the first identity. By the definition of $\mathcal{C}(T)$, for $0_{G^{n-1}} \neq \rho \notin \mathcal{C}(T)$, either

α) exists $l \notin L_e$, $l \neq R$ with $\rho_l \neq 0_G$, or

β) exist $l, j \in L_e$, such that $\rho_l \neq \rho_j$.

In α), take an $\eta \in \hat{G}$ such that $\eta(\rho_l) \neq 1$. Such an η exists, since by Lemma 1(ii) the matrix $[\chi(g)]$ is regular, and it already has a column full of 1's, namely, for $\rho = 0$. In (6),

assign to the character $\chi = (\chi_1, ..., \chi_l, ..., \chi_{n-1})$ the character $\chi = (\chi_1, ..., \eta + \chi_l, ..., \chi_{n-1})$. Observe that on the one hand we just permuted the terms in the sum (6), and therefore fixed the value of the sum; on the other hand, we multiplied the sum by $\eta(\rho_l) \neq 1$. Hence, the sum is 0.

In β), take an $\eta \in \hat{G}$ such that $\eta(\rho_j - \rho_l) = \eta(\rho_j)\eta^{-1}(\rho_l) \neq 1$. Such an η exists, since like in α), $\rho_j - \rho_l$ would yield a second column full of 1's in $[\chi(g)]$, contradicting the regularity. In (6), assign to the character $\chi = (\chi_1, ..., \chi_l, ..., \chi_j, ..., \chi_{n-1})$ the character $\chi = (\chi_1, ..., \chi_l - \eta, ..., \chi_j + \eta, ..., \chi_{n-1})$. Observe that on the one hand we just permuted the terms in the sum (6), and therefore fixed the value of the sum; on the other hand, we multiplied the sum by $\eta(\rho_j - \rho_l) \neq 1$. Hence, the sum is 0.

The proof of the conjugated exponent version is virtually the same and we leave it to the reader. ∎

We give an alternative logarithmic formulation of Theorem 5, since this logarithmic formulation was discovered and published for $G = Z_2$ [H] and $G = Z_2 \times Z_2$ [SHSE]. Let $K = [h(g)]$ denote the matrix, in which rows correspond to $h \in \hat{G}$ and columns correspond to $g \in G$; let $H = [\chi(\sigma)]$ denote the matrix, in which rows correspond to $\chi \in \hat{G}^{n-1}$ and columns correspond to $\sigma \in G^{n-1}$. Let the logarithm of a vector denote the vector of logarithms of the components. Let $\mathbf{f}$ denote the vector of f_σ's ($\sigma \in G^{n-1}$), and let $\mathbf{p}_e$ denote the vector of $p_e(g)$'s ($g \in G$) for every $e \in E(T)$.

Theorem 6.

$$[H^{-1} \log H\mathbf{f}]_\rho = \begin{cases} 0, & \text{if } 0 \neq \rho \notin C(T), \\ [K^{-1} \log K\mathbf{p}_e]_h, & \text{if } \rho = \rho^{e,h} \in C(T), \\ \sum_{e \in E(T)} \sum_{h \in G} [K^{-1} \log K\mathbf{p}_e]_h, & \text{if } \rho = 0. \end{cases} \quad (7)$$

Proof. Take the logarithm of the conjugated exponent versions of the identities in Theorem 5, and use the identities for the adjugates

$$\frac{1}{|G|} K^* = K^{-1} \quad \text{and} \quad \frac{1}{|G|^{n-1}} H^* = H^{-1}$$

to get rid of the powers of group orders. ∎

4. Series expansion

We say that a vector $\mathbf{x}$ of x_σ's ($\sigma \in G^{n-1}$) is *regular*, if $\sum_\sigma x_\sigma = 1$, x_σ is non-negative real, $x_0 > 1/2$. For the expansions in this Section regularity is a convenient sufficient condition, although it is not necessary.

Theorem 7. *For a regular* $\mathbf{x}$ *and* $\sigma \neq 0$,

$$[H^{-1}\log H\mathbf{x}]_\sigma = \sum_{r=1}^{\infty} \frac{(-1)^{r+1}}{r} \sum_{\substack{(\sigma_1,\ldots,\sigma_r):\\ \sigma_1+\ldots+\sigma_r=-\sigma\\ \sigma_i\neq 0}} \prod_{i=1}^{r} \frac{x_{\sigma_i}}{x_0}.$$

Proof. We use regularity to establish

$$|\sum_{\sigma:\sigma\neq 0} \chi(\sigma)x_\sigma| < x_0. \tag{8}$$

Indeed,

$$|\sum_{\sigma:\sigma\neq 0} \chi(\sigma)x_\sigma| \leq \sum_{\sigma:\sigma\neq 0} |\chi(\sigma)||x_\sigma| = \sum_{\sigma:\sigma\neq 0} x_\sigma = 1 - x_0 < x_0.$$

We start with

$$[H\mathbf{x}]_\chi = \sum_\sigma \chi(\sigma)x_\sigma = x_0(1 + \sum_{\sigma:\sigma\neq 0} \chi(\sigma)\frac{x_\sigma}{x_0}).$$

We combine (8) with the fact that radius of convergence of the Taylor series of $\log z$ at $z = 1$ is 1:

$$[\log H\mathbf{x}]_\chi = \log x_0 - \sum_{r=1}^{\infty} \frac{(-1)^{r+1}}{r} (\sum_{\sigma:\sigma\neq 0} \chi(\sigma)\frac{x_\sigma}{x_0})^r.$$

Hence

$$[H^{-1}\log H\mathbf{x}]_\rho = \sum_{r=1}^{\infty} \frac{(-1)^{r+1}}{r} \sum_\chi \chi(\rho)(\sum_{\sigma_1:\sigma_1\neq 0} \chi(\sigma_1)\frac{x_{\sigma_1}}{x_0})\cdots(\sum_{\sigma_r:\sigma_r\neq 0} \chi(\sigma_r)\frac{x_{\sigma_r}}{x_0})$$

$$= \sum_{r=1}^{\infty} \frac{(-1)^{r+1}}{r} \sum_{\substack{(\sigma_1,\ldots,\sigma_r):\\ \sigma_i\neq 0}} \frac{x_{\sigma_1}x_{\sigma_2}\ldots x_{\sigma_r}}{x_0^r} \sum_\chi \chi(\rho+\sigma_1+\ldots+\sigma_r).$$

Now observe that $\sum_\chi \chi(\rho+\sigma_1+\ldots+\sigma_r)$ vanishes, except if $\rho+\sigma_1+\ldots+\sigma_r = 0$ according to the summation in the theorem; and in this case its value is $|G|^{n-1}$.∎

Corollary 8. *For a regular* $\mathbf{x}$ *and* $\sigma \neq 0$, *we have the first and second order approximations*

$$[H^{-1}\log H\mathbf{x}]_\sigma \approx \frac{x_\sigma}{x_0},$$

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

respectively.∎

Let p^{*k} denote the k-order convolution of the transition mechanism with itself as defined in Theorem 4; now Theorem 4 and a standard inclusion-exclusion argument allows for the following expansion:

Corollary 9. *For regular* $\mathbf{f}$ *and* $\sigma \neq 0$,

$$[H^{-1}\log H\mathbf{f}]_\sigma = \sum_{r=1}^{\infty}\sum_{k=1}^{r}\frac{(-1)^{k+1}\binom{r}{k}f_{-\sigma}(T,p^{*k})}{rf_0^k(T,p)}.\blacksquare$$

5. Invariants

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. 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 $0 \neq \rho \notin C(T)$, on the basis of Theorem 5, we term the δ_ρ's as the *canonical invariants* of the tree T.

Now we are ready to state the main results of this Section; writing $\mathbf{p}_e$ in vector form we put $p_e(0)$ into the first coordinate.

Theorem 10. *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.*

Proof. (i) Take an $e \in E(T)$ such that $L_e \notin Split(T')$. Then $\rho^{e,h} \notin C(T')$ for $0 \neq h \in G$; and the hypothesis of (i) implies $[H^{-1} \log H\mathbf{f}]_{\rho^{e,h}} = 0$ for all $h \neq 0$. On the other hand, (7) implies $[H^{-1} \log H\mathbf{f}]_{\rho^{e,h}} = [K^{-1} \log K\mathbf{p}_e]_h$ for all $h \neq 0$. Hence, $[K^{-1} \log K\mathbf{p}_e]_h = 0$ for all $h \neq 0$. In other words, $K^{-1} \log K\mathbf{p}_e = (x, 0, ..., 0)^T$ for some number x, and hence $\log K\mathbf{p}_e = (x, x, ..., x)^T$, $K\mathbf{p}_e = (exp(x), exp(x), ..., exp(x))^T$, and finally $\mathbf{p}_e = (exp(x), 0, ..., 0)$, i.e. the edge e must have been passive.

(ii) is a simple application of (i). Observe that the hypothesis of (iii) implies the hypothesis of (ii), and hence the conclusion of (ii) holds.

We finish the proof by (iv). We prove more: the δ_ρ's are algebraically independent for $\rho \in G^{n-1}$. By the multivariate Taylor formula the δ_ρ's are algebraically independent iff the $\delta_\rho + 1$'s are. Suppose that

$$\sum_s \lambda_s \prod_{\rho \in G^{n-1}} (\delta_\rho + 1)^{i_{\rho,s}} = \sum_s \lambda_s \prod_{\chi \in \hat{G}^{n-1}} R_\chi^{\sum_\rho i_{\rho,s} \overline{\chi(\rho)}} \tag{9}$$

is identically zero in a neighborhood of $x_0 = 1$, $x_\sigma = 0$ with a certain finite set of complex coefficients λ_s and non-negative integer exponents $i_{\rho,s}$. We may assume without loss of generality that $s \neq s'$ implies that for some ρ we have $i_{\rho,s} \neq i_{\rho,s'}$. Since the invertible linear transformation H turns the x_σ's into the R_χ's, we may study the vanishing of (9) in the independent variables R_χ's, all in a neighborhood of 1. Having independent variables, the only way of vanishing (9) is cancellation, i.e. for some $s \neq s'$ and all $\chi \in G^{n-1}$

$$\sum_{\rho \in G^{n-1}} i_{\rho,s} \overline{\chi(\rho)} = \sum_{\rho \in G^{n-1}} i_{\rho,s'} \overline{\chi(\rho)}. \tag{10}$$

The matrix H and its conjugate $\bar{H}$ are regular; hence (10) implies $i_{\rho,s} = i_{\rho,s'}$ for all $\rho \in G^{n-1}$, a contradiction. ∎

The reader might ask if logarithms and all the resulting fuss about smallness of some quantities are necessary to obtain our results. Therefore we show a simple example to point

out that Theorem 10(*iii*) turns into false if we drop these conditions. Take an arbitrary tree T and define the transition mechanism by $p_e(g) = 1/|G|$ for all $e \in E(T)$, $g \in G$. Clearly, f_σ will follow the uniform distribution independently of the topology of the tree, contrary to Theorem 10(*iii*).

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

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

Clearly, we obtained polynomial invariants, of which most of Theorem 10 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 [ES] 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 $2^{m(n-1)} - |\mathcal{C}(T)| - 1 = 2^{m(n-1)} - |E(T)|(2^m - 1) - 1$; 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. Our Theorem 10(*i*) is some support for the conjecture of Evans and Speed.

6. Kimura's models of molecular evolution

One assumes that the process of evolution is described by a tree. In this tree the labelled leaves denote some existing species represented by corresponding segments of aligned DNA sequences, the unlabelled branching vertices may denote unknown extinct ancestors. Let r denote the immediate ancestor of the closest common ancestor of a given set of existing species. We define the *true tree* of this set of species by taking the subtree induced by them and r in the tree describing the process of evolution and undoing the vertices of degree two.

The very problem of reconstruction may be put in this way: given a set of species with corresponding segments of aligned DNA sequences, find the true tree.

For $G = Z_2$, the model described in Section 2 specializes to a model of Cavender [C], 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 purine-pyrimidine alphabet [H], [HP1], [HP2]. Our part is 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; see [SHSE]. However, it is theoretically possible to apply our calculus to either of the two Abelian groups of order 20 (if the transition mechanisms of amino acids follow either of these groups), and also to Z_4, in Kimura's 2-parameter model and the Jukes-Cantor model (see below). We explain the $G = Z_2 \times Z_2$ case in detail, the explanation also applies, mutatis mutandis, to $G = Z_2$.

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 [ES]. Identify the elements of $Z_2 \times Z_2$ with the four nucleotides, such that A is the unity. 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.

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. 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. 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 biologists easily recover it by a method called *outgroup comparison.* The root that we use, is, like in Section 2, *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 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.

We had a set of species with corresponding segments of aligned DNA sequences. We selected an arbitrary species for R and we standardized the sequences from R, and obtained an f'_σ relative frequency of the colouration σ among the bits. Now we face the following problem: which tree T and transition mechanism p yield leaf colouration probabilities $f_\sigma = f'_\sigma$ for all σ? Working with real data, we must be satisfied with the

best approximation in a reasonable norm. Having the transition mechanism of the true tree allows for estimating a time scale, i.e. how far ago in time the evolutionary events in question did happen. We note here, that the model of Section 2 does not imply the existence of the logarithms; however, for real data, there is no problem with them, due to the empirical fact that $f'_0 > 1/2$. Working with $\mathbf{f}$ arising from the model of Section 2, Theorem 6 tells the edges of the tree, and one can obtain the transition mechanism, i.e. $\mathbf{p}_e$ for all edges as well. The message of Theorem 10(*iii*) is, that we may expect a *unique* tree to yield the observed relative frequencies of leaf colourations.

Working with empirical $\mathbf{f}'$, the closest tree method [H], which is a branch-and-bound algorithm, determines then the evolutionary tree and its transition mechanism, 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 significance of the series expansion is that a second order approximation of $H^{-1}\log H\mathbf{f}'$ can be computed $O(t^2)$ time, where t is the number of nonzero f'_σ's, which is subexponential by our experience for real data. The use of the second order approximation is expected to be superior to computing of $H^{-1}\log H\mathbf{f}'$ by Fast Fourier Transform on real data; this is still to be tested.

The great advantage of using invariants is that one may discriminate against some trees without (strong) assumptions regarding the transition mechanism. Invariants were introduced by Cavender and Felsenstein [CF], [C2], [C3] and Lake [L]; and recently Evans and Speed [ES] gave an algebraic procedure based on Fourier analysis to decide if a polynomial is invariant or not for $G = Z_2^m$. The literature shows that all the efforts went for polynomial invariants. There is a good reason to look for linear invariants, namely, they are subject to reliable statistical methods. However, there are cases, when linear invariants are known not to exist, including Kimura's 3-parameter model [ES]. In lack of linear invariants, there is at most a theoretical reason to prefer polynomial invariants.

The advantage of our canonical invariants to other invariants is, that they come from a predetermined list, and if you need the canonical invariants of a tree, you just pick the right elements from the list. If it comes to application of our polynomial invariants, then values of the polynomial functions must be computed instead of the polynomials, since computer algebra in many variables is rather prohibitive.

We see the significance of the Fourier calculus on evolutionary trees in the fact, that it puts the tree reconstruction to the basis of the generally accepted theory of molecular

evolution by Kimura, while most tree reconstruction techniques lack any such mechanism in the background.

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.

[ES] S. N. Evans, T. P. Speed, Invariants of some probability models used in phylogenetic inference, *Annals of Statistics*, in press.

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

[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 spectral analysis for evolution, submitted to *Proc. Natl. Acad. Sci.*

[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.

[Kö] T. W. Körner, *Fourier Analysis*, Cambridge University Press, Cambridge, 1988.

[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.

[S] M. A. Steel, *Distributions on Bicoloured Evolutionary Trees*, Ph. D. Thesis, Massey University, Palmerston North, 1989.

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

[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, in press.

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

L. A. Székely, Department of Computer Science, Eötvös University, H-1088 Budapest, and

Department of Mathematics, University of New Mexico, Albuquerque, NM 87131, U.S.A.

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

P. L. Erdős, Hungarian Academy of Sciences, H-1055 Budapest