Appl. Math. Lett. Vol. 6, No. 2, pp. 13–16, 1993
Printed in Great Britain.

0893-9659/93 $6.00 + 0.00

A FOURIER INVERSION FORMULA FOR EVOLUTIONARY TREES

L. A. SZÉKELY*
Department of Computer Science, Eötvös University, Budapest, Hungary
and Institut für diskrete Mathematik, Bonn, Germany
P. L. ERDŐS
Department of Mathematics, University of Groningen, Groningen, The Netherlands
and Hungarian Academy of Sciences, Budapest, Hungary
M. A. STEEL*
Department of Mathematics, Massey University, Palmerston North, New Zealand
and Universität Bielefeld, Zentrum für interdisziplinäre Forschung, Bielefeld, Germany
D. PENNY
Department of Zoology, Massey University, Palmerston North, New Zealand

(Received and accepted June 1992)

Abstract—We establish a pair of identities, which will provide a useful tool in the reconstruction of evolutionary trees in Kimura's 3-parameter model.

The starting point of this paper was an attempt for a better understanding and generalization of an Hadamard inverse pair of formulae, which was used in statistics by Cooper [1], in image processing by Andrews [2, Chapters 6,7], and in information theory by Whelchel and Guinn [3]. Recently, Hendy and Penny [4] applied this technique to the spectral analysis of phylogenetic data, for two-state character sequences, and they asked for generalization to four-state character sequences, which is the form of a nucleotide sequence. The most invaluable tool for our work was [5], where discrete Fourier analysis is applied to somewhat similar problems. Let us start with the original theorem of Hendy and Penny [6].

Suppose we are given a tree T with leaf set L, out of which one is a root called R. Set $|L| = n$. Toss a coin independently for every edge e of the tree with probability p_e for a head and probability $1 - p_e$ for a tail. Let σ denote a subset of $L \setminus \{R\}$. Let f_σ denote the probability that σ is precisely the set of leaves for which the unique Rl path contains an odd number of edges with head outcome.

Let X be a subset of L of even cardinality. Match the vertices of X somehow with paths in the tree and notice that the set of edges $P(T, X)$ used by an odd number of paths is independent of the matching. Set

$$r_X = \prod_{e \in P(T,X)} (1 - 2p_e).$$

Then, from [6] one has the following formulae:

$$r_X = \sum_{\sigma \subseteq L \setminus \{R\}} (-1)^{|\sigma \cap X|} f_\sigma, \qquad f_\sigma = \frac{1}{2^{n-1}} \sum_{X : |X| \text{even}} (-1)^{|\sigma \cap X|} r_X.$$

This paper was written while the third author was visited by the other three authors at the Zentrum für interdisziplinäre Forschung, Bielefeld.
*Research supported by Alexander v. Humboldt-Stiftung.

We give a sought-for generalization of these formulae, in which, in addition, independence is an unnecessary assumption, and we also provide a new proof for the original theorem.

Let us continue with a tree T with leaf set L, out of which one is a root called R. Set $|L| = n$, let m be an arbitrary positive integer. Toss a coin for every edge e of the tree and number i $(1 \leq i \leq m)$ with probability p_e^i for a head and probability $1 - p_e^i$ for a tail. (Now independence is not assumed, neither on the same edge, nor on different edges.) Let $\sigma_1, \sigma_2, \ldots, \sigma_m$ denote an ordered m-tuple of subsets of $L \setminus \{R\}$. Let $f_{\sigma_1,\sigma_2,\ldots,\sigma_m}$ denote the probability that $\forall i (l \in \sigma_i \iff$ the unique Rl path contains an odd number of edges with head outcome out of the i-type experiments over the edges of that path).

Let $X_1, X_2, \ldots, X_m$ be an ordered m-tuple of subsets of L, where all subsets are of even cardinality. Match the vertices of X_i somehow with paths in the tree and notice that the set of edges $P(T, X_i)$ used by an odd number of paths is independent of the matching. Define $\epsilon_{ie} = 1$, if the outcome of the i^{th} toss over the edge e is head, and $= 0$, if the outcome is tail. Set

$$\mathcal{E} = \{\epsilon_i(e) : i \in \{1, 2, \ldots, m\}, e \in P(T, X_i)\},$$

a set of summation variables, which will take 0 and 1 values; and set

$$r_{X_1,\ldots,X_m} = \cdots \sum_{\epsilon_i(e)=0}^{1} \ldots (-1)^{\sum_{i=1}^{m} \sum_{e \in P(T,X_i)} \epsilon_i(e)} P(\forall i, e : \ e \in P(T, X_i) \Longrightarrow \epsilon_{ie} = \epsilon_i(e)),$$

where the summation is over all elements of $\mathcal{E}$. Now we have the following theorem.

THEOREM.

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

$$f_{\sigma_1,\ldots,\sigma_m} = \frac{1}{2^{m(n-1)}} \sum_{X_1,\ldots,X_m} (-1)^{\sum_{i=1}^{m} |\sigma_i \cap X_i|} r_{X_1,\ldots,X_m}.$$

PROOF. We need two well-known lemmas to complete the proof.

LEMMA 1. *(Rényi) Let $A_1, \ldots, A_n$ be any events, $B_i = f_i(A_1, \ldots, A_n)$ $(i = 1, 2, \ldots, k)$ Boolean polynomials in $A_1, \ldots, A_n$ and $c_1, \ldots, c_k$ reals. Then $\sum_{i=1}^{k} c_i \, Prob(B_i) \geq 0$ holds for every $A_1, \ldots, A_n$, provided it holds in those cases when $Prob(A_j) = 0$ or 1 for $j = 1, 2, \ldots, n$.*

PROOF. See [7; 8, p. 20].

LEMMA 2. *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 $g \in G$,

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

PROOF. See [5] or any relevant textbook.

Let G consist of the ordered m-tuples $(\sigma_1, \ldots, \sigma_m)$ with the operation

$$(\sigma_1, \ldots, \sigma_m) + (\rho_1, \ldots, \rho_m) = (\sigma_1 \Delta \rho_1, \ldots, \sigma_m \Delta \rho_m),$$

where Δ denotes symmetric difference of sets, and let $\hat{G}$ consist of the ordered m-tuples $(X_1, \ldots, X_m)$ (all of them are of even size), while the operation is

$$(X_1, \ldots, X_m) + (Y_1, \ldots, Y_m) = (X_1 \Delta Y_1, \ldots, X_m \Delta Y_m).$$

It is easy to see that for an arbitrary fixed element $(X_1, \ldots, X_m)$ of $\hat{G}$, the operation

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

is a homomorphism into the multiplicative group of $\{1, -1\}$; therefore, the elements of $\hat{G}$ are all characters. Since we have the right number of them, $\hat{G}$ is the character group. Hence, Lemma 2(ii) applies, and it is sufficient to prove the first formula of the theorem.

By Lemma 1, it is sufficient to prove the first formula for $p_f^j = 0$ or 1. In this case, both of the RHS of the first formula and the definition of $r_{X_1,\ldots,X_m}$ will contain only one nonzero term, namely a ± 1. The sign will be correct if

$$(-1)^{\sum_{i=1}^m \sum_{e \in P(T,X_i)} \epsilon_{ie}} = (-1)^{\sum_{i=1}^m |\sigma_i \cap X_i|}.$$

We claim even more the following lemma.

LEMMA 3. *For all j we have* $\sum_{e \in P(T,X_i)} \epsilon_{ie} \equiv |\sigma_i \cap X_i| \pmod 2$.

PROOF. It could even suffice to say at this point, that Lemma 3 is implied by the truth of the two-state character formula, but it is instructive to see the straightforward proof. LHS $\equiv$ sum of the number of i-type heads on paths that make a matching of the vertices of $X_i \equiv$ sum of the number of i-type heads on the Rl-paths, $l \in X_i \equiv$ RHS (mod 2).

We now make some comments regarding the proof. Since the Kronecker product of Hadamard matrices is an Hadamard matrix, it was not necessary to use Lemma 2, since we may have used the inverse of an Hadamard matrix. However, it gave extra insight, which was necessary to find the first formula.

If all the tosses are independent, then

$$r_{X_1,\ldots,X_m} = r_{X_1} r_{X_2} \cdots r_{X_m}, \quad \text{and} \quad f_{\sigma_1,\ldots,\sigma_m} = f_{\sigma_1} f_{\sigma_2} \cdots f_{\sigma_m},$$

and our theorem traces back to the theorem of [6] by multiplication. This was the way to find the theorem. If $m = 1$ and the tosses are independent, then we get back the theorem of [6] after some algebra.

The practical importance of the theorem is particular for $m = 2$. We follow [5]. Take the Kleinian group $Z_2 \times Z_2$ with generators a and b, and assign an element of the Kleinian group to every edge of the tree in the following way: the first head means a, the second means b, tail means unity, and in this way multiply together the outcomes of the two tosses. Assign to the root the unity, and to every other vertex assign the product of group elements along the unique path connecting the root to that particular vertex. We have a random 4-colouration of the vertices of the tree, and if we assume the independence of tosses performed on distinct edges, then this is the Kimura 3-parameter model of evolutionary trees in the group theoretical setting of [5]. Just make a correspondence like $A \leftrightarrow e$ (unity), $G \leftrightarrow a$, $C \leftrightarrow b$, $T \leftrightarrow ab$.

Kimura's 3-parameter model does not allow the independence of the two tosses associated with the same edge. Therefore, all the tricks we did were necessary. The applications require the fast computation of r_{X_1,X_2}. For Kimura's 3-parameter model, the following identity helps the fast computation:

$$r_{X_1,X_2} = \prod_{e \in P(T,X_i)} (1 - 2P(\epsilon_{ie} = 1)) \prod_{e \in P(T,X_1) \cap P(T,X_2)} (P(\epsilon_{1e} = \epsilon_{2e}) - P(\epsilon_{1e} = 1 - \epsilon_{2e})),$$

where the first product is over those edges e, which belong to just one $P(T, X_j)$. This identity, combined with the theorem, allows for one to recover from the f_{σ_1,σ_2}'s the four probabilities

associated with each edge of the tree under the Kimura 3-parameter model. This opens the way to applications to the reconstruction of trees from genetic sequences, by generalizing the "closest tree" approach of Hendy and Penny [6] and the subsequent "spectral analysis" of phylogenetic data [4]. The actual applications will appear elsewhere.

Finally, we note that our method gives a pair of inverse formulae in a more general setting, where $r(X_1, \ldots, X_m)$ and $f(\sigma_1, \ldots, \sigma_m)$ are arbitrary complex-valued functions; namely, if one of our two formulae holds, then the other holds as well.

REFERENCES

1. B.E. Cooper, The extension of Yates' 2^n algorithm to any complete factorial experiment, *Technometrics* **10**, 575–577 (1968).
2. H.C. Andrews, *Computer Techniques in Image Processing*, Academic Press, New York, (1970).
3. J.E. Whelchel and D.F. Guinn, The fast Fourier-Hadamard transform and its use in signal representation and classification, In *Eascon 1968 Convention Record*, pp. 561–573.
4. M.D. Hendy and D. Penny, Spectral analysis of phylogenetic data, (preprint), University of Bielefeld, ZiF-Nr. 91/23, (1991).
5. S.N. Evans and T.P. Speed, Invariants of some probability models used in phylogenetic inference, *Annals of Statistics* (to appear).
6. M.D. Hendy and D. Penny, A framework for the quantitative study of evolutionary trees, *Systematic Zoology* **38** (4), 297–309 (1989).
7. A. Rényi, *Foundations of Probability*, Holden Day, San Francisco, (1970).
8. L. Lovász, *Combinatorial Problems and Exercises*, North-Holland, Amsterdam, (1979).