Title
Ubiquity of synonymity: almost all large binary trees are not uniquely identified by their spectra or their immanantal polynomials

Journal
Algorithms for Molecular Biology, 7(1)

ISSN
1748-7188

Authors
Matsen, Frederick A
Evans, Steven N

Publication Date
2012-05-21

DOI
http://dx.doi.org/10.1186/1748-7188-7-14

Peer reviewed

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Ubiquity of synonymity: almost all large binary trees are not uniquely identified by their spectra or their immanantal polynomials

Frederick A Matsen [1*] and Steven N Evans[2]

Abstract

Background: There are several common ways to encode a tree as a matrix, such as the adjacency matrix, the Laplacian matrix (that is, the infinitesimal generator of the natural random walk), and the matrix of pairwise distances between leaves. Such representations involve a specific labeling of the vertices or at least the leaves, and so it is natural to attempt to identify trees by some feature of the associated matrices that is invariant under relabeling. An obvious candidate is the spectrum of eigenvalues (or, equivalently, the characteristic polynomial).

Results: We show for any of these choices of matrix that the fraction of binary trees with a unique spectrum goes to zero as the number of leaves goes to infinity. We investigate the rate of convergence of the above fraction to zero using numerical methods. For the adjacency and Laplacian matrices, we show that the *a priori* more informative immanantal polynomials have no greater power to distinguish between trees.

Conclusion: Our results show that a generic large binary tree is highly unlikely to be identified uniquely by common spectral invariants.

Background

Tree shape theory furnishes numerical statistics about the structure of a tree [1,2]. (Because we are interested in applications of tree statistics to trees that describe the structure of branching events in evolutionary histories, we will, for convenience, always take the term *tree* without any qualifiers to mean a **rooted, binary tree without any branch length information or labeling of the vertices**.) Such statistics have two related uses. Firstly, they can be used in an attempt to tell whether two trees are actually the same and, secondly, they can be used to indicate the degree of similarity between two trees with respect to some criterion.

Examples of the latter use are the testing of hypotheses about macroevolutionary processes and the detection of bias in phylogenetic reconstruction. Historically, numerical statistics for such purposes have attempted to capture the notion of the *balance* of a tree, which is the degree to which daughter subtrees are the same size. The balance is typically measured by ad-hoc formulae that are often selected for statistical power to distinguish between two different distributions on trees [3,4]. In previous work we investigated the possibility of describing the shape of the tree using a list of numbers rather than just a single number [5,6].

A mathematically "canonical" approach to finding a list of such numbers is to use information derived from matrix representations of the trees. We first describe the matrix representations of a tree that we will consider.

In algebraic graph theory [7], the basic matrix associated to a graph is the *adjacency matrix* $A(G)$, whose ij^{th} entry is one if i and j are connected by an edge, and zero otherwise. From a probabilistic point of view, the more natural matrix to associate with a graph is the *Laplacian matrix* $L(G)$, which is the infinitesimal generator of the natural random walk on the graph and is given by $A(G) - D(G)$, where $D(G)$ is the diagonal matrix of vertex degrees. It is clear that a graph can be recovered from either its adjacency of Laplacian matrix. Some authors, such as [8], define the Laplacian to be $D(G)^{-1/2}L(G)D(G)^{1/2}$. Note that this difference is not relevant if one is only considering

* Correspondence: matsen@fhcrc.org
[1]Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
Full list of author information is available at the end of the article

characteristics of the matrix L such as eigenvalues that are invariant under similarity transformations.

Readers in the phylogenetics community may be more familiar with the *pairwise distance matrix* [1,9]. The distance matrix P given a leaf-labeling 1, . . ., n has as its ij^{th} entry the length of the path between leaf i and leaf j. Any leaf-labeled tree is uniquely determined by its distance matrix. These matrices have also been extensively studied as discrete metric spaces [10,11].

The definition of the adjacency and Laplacian matrices requires a numbering of the vertices, while the definition of the distance matrix requires a numbering of the leaves. Because we are considering unlabeled trees (that is, we identify trees that are equivalent in the usual sense of graph-theoretic isomorphism), we are only interested in tree statistics that are invariant under renumbering. Algebraically, this means that we are only interested in features of the associated matrix that are unaffected by similarity transformations via a permutation matrix. The most obvious such statistics are the eigenvalues.

The adjacency and Laplacian matrices and their eigenvalues are familiar objects in the area of spectral graph theory [7,8,12]. The eigenvalues of the adjacency matrix tend to contain combinatorial information about the graph, such as bounds on the chromatic number. The eigenvalues of the Laplacian give information of a more geometric flavor, such as the equivalent of the surface area to volume ratio of subgraphs of a graph. As well as having connections to the theory of random walks on graphs, the Laplacian eigenvalues can be used to define the expander graphs, an important class of graphs that have applications in coding theory. Therefore, it would not be too surprising if the these eigenvalues were a convenient way to summarize information about a tree, thus giving a nice collection of tree statistics.

Similarly, it seems plausible that the eigenvalues of the pairwise distance matrix could contain quite a lot of information about the tree that could be used to compare trees. Moreover, although the distance matrix formally contains the same information as the adjacency or Laplacian matrices, the transformation that takes the distance matrix to one of the other two is distinctly non-linear, and hence there is no reason to believe that there is any simple connection between the corresponding eigenvalues.

We demonstrate below that not only do there exist pairs of trees that have the same spectrum as another tree for the adjacency, Laplacian, and distance matrices, but that this is the rule rather than the exception as the trees become large, in the sense that the fraction of trees with a given number of leaves that have a unique adjacency, Laplacian, or distance spectrum goes to zero as the number of leaves goes to infinity.

The basic methodology that we use to prove this result was first established in [13] and developed in [14] for general (that is, not necessarily bifurcating) graph-theoretic trees in the case of the adjacency and Laplacian matrices. The present paper provides the first results of this type concerning rooted bifurcating trees, as well as the first examination of such results for the pairwise distance matrix. The key idea is to establish that certain pairs of trees T_1 and T_2 have the following *exchange property* for a given matrix representation: that exchanging T_1 for T_2 as subtrees of a given tree does not change the spectrum for that matrix representation. This is a stricter requirement than simply having the same spectrum (Figure 1). It then becomes a matter of showing that the number of trees with a given number of leaves is asymptotically of larger order than the number of trees with the same number of leaves that don't have a particular subtree. For this we build on the generating function argument used in [15] for asymptotic estimates of the number of unlabeled rooted bifurcating trees (see the section *Asymptotic numbers of trees*).

One possible explanation for this phenomenon is that two diagonalizable matrices have the same spectrum if they are similar via an arbitrary similarity transformation rather than just via a permutation transformation, and this suggests considering features of a matrix that are invariant under permutation similarities but not more general ones. We will now describe a feature of a matrix, its *immanantal polynomial,* that has this property.

The *immanant* is a generalization of the determinant. Recall that the determinant of a matrix $A = (a_{ij})$ is given by

$$\det(A) := \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_i a_{i\sigma(i)},$$

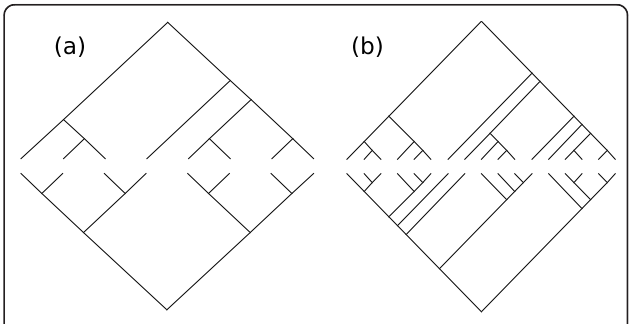

Figure 1 Pairs of trees with similar algebraic properties. Figure (a) shows the smallest pair of rooted binary trees with the same adjacency and Laplacian spectrum; these trees do not have the corresponding exchange property. Figure (b) shows two trees with the exchange property for the adjacency, Laplacian, and pairwise distance matrices.

where the sum is over the symmetric group of permutations of {1, 2,..., n} and sgn(σ) is the *sign* of the permutation σ.

The function sgn is a particular example of a *character* of an *irreducible representation* of the symmetric group. It would take us too far afield to define these notions here, but excellent treatments may be found in [16-19]. We note, however, that the irreducible characters are constant on the conjugacy classes of the symmetric group (recall that two permutations belong to the same conjugacy class if and only if they have the same cycle structure) and they form a basis for the vector space of functions with this property (the *class functions)*.

Our use of characters is simply to define the immanant

$$\mathcal{I}_\chi(A) := \sum_{\sigma \in S_n} \chi(\sigma) \prod_i a_{i\sigma(i)}$$

of a matrix A for the irreducible character χ. A discussion of immanants may be found in [20,21]. The *immanantal polynomial* for a character χ of a matrix is the corresponding generalization of the characteristic polynomial; that is, it is the polynomial $x \mapsto \mathcal{I}_\chi(xI - A)$. Because the characters are class functions, an immanantal polynomial is invariant under similarity by permutation matrices, but it will not typically be invariant under more general similarities.

Unfortunately, as we show in Lemma 2, for either the adjacency or Laplacian matrix the following two conditions on a pair of trees are equivalent:

- the spectra are equal,
- the immanantal polynomials are equal for all irreducible characters.

Consequently, the immanantal polynomials for the adjacency and Laplacian matrices provide no more distinguishing power than the spectra and, in particular, a vanishing fraction of large trees have a unique set of immanantal polynomials for these matrices. We do not know if the same fact is true for the immanantal polynomials of the distance matrix.

Our main result is thus the following.

Theorem 1. *Let t_n be the number of trees with n leaves. For either the adjacency, Laplacian, or pairwise distance matrix, let l_n be the number of trees with n leaves that do not share their spectrum with another tree. Then, the fraction l_n/t_n goes to zero as n goes to infinity. For the adjacency and Laplacian matrices, the same result holds if we replace the spectrum by the complete set of immanantal polynomials.*

The rate of convergence of the fraction in Theorem 1 is also of interest. If it is extremely slow then the existence of trees with shared spectra may not be practically relevant for the construction of informative tree shape statistics. We investigate this matter numerically towards the end of the paper.

Several other research groups have investigated problems that are related to, but different from, those investigated here. Steyaert and Flajolet [22] investigate the occurrences of subtrees in the case of "planar" trees, i.e. trees that are equipped with an order of subtrees at every internal node. These planar trees are substantially easier to analyze: for example, there is a nice closed form generating function for the numbers of such trees (the Catalan numbers). In contrast, the generating function for the number of trees considered here is only given as the solution of a functional equation and there is no closed form expression for the numbers of such trees. Graham and Lovasz [23] investigate the spectra of distance matrices of trees, but their distance matrices are defined in terms of vertex-to-vertex distances, rather than the leaf-to-leaf definition. The leaf-to-leaf distance matrix is a principal sub-matrix of the vertex-to-vertex one, and there are interlacing relations between the spectrum of a matrix and one of its principal submatrices (2). However it is not a priori the case that if two matrices have the same spectrum, then the principal submatrices with the same rows and columns will also have the same spectrum. In a similar vein, the vertex-to-vertex distance matrix can be constructed from the leaf-to-leaf distance, but the construction involves considering whether certain linear inequalities hold and so it isn't a procedure that will, a priori, transform spectra in a simple way.

More recently, Bhamidi, Evans, and Sen [24] have proven that, subject to weak general conditions, many ensembles of random trees have the property that, with probability converging to one as the number of leaves goes to infinity, a realization shares its spectrum with another tree. Their conditions are easiest to check when the ensemble can be embedded in a general continuous-time branching process where individuals give birth to a possibly random number of offspring at the arrival times of a point process up to a possibly infinite death time, and those offspring go on to behave as independent copies of their parent. This particular framework covers examples such as random recursive trees, linear preferential attachment trees, uniform rooted unordered labeled trees, and Yule trees. However, we have been unable to embed the ensemble considered here in a suitable continuous-time branching process or otherwise check the general conditions of [24].

The computer code used in this paper was written in `OCaml` [25] and has been made available at http://github.com/matsen/ubiquity_synonymity, along with the results produced by this code.

Algebraic preliminaries concerning spectra and immanantal polynomials

Equality of adjacency and Laplacian spectra implies equality of immanantal polynomials

In order to prove results for the adjacency and Laplacian matrices simultaneously, we define for a tree T and arbitrary real numbers y and z the *generalized Laplacian* $\tilde{L}(T) := yD(T) + zA(T)$ (recall that $A(T)$ is the adjacency matrix and $D(T)$ is the diagonal matrix of vertex degrees). We define the corresponding *generalized Laplacian immanantal polynomial* of the tree T with r vertices to be

$$x \mapsto \mathcal{I}_\chi\left(xI - \tilde{L}(T)\right)$$

for an irreducible character χ of the symmetric group S_r.

The generalized Laplacian immanantal polynomial can be computed in a simple combinatorial fashion as follows. Define a *k-matching* to be a set of k pairwise disjoint edges of the tree (that is, a set of edges such that no two share a common vertex). Let $M_k(T)$ denote the set of k-matchings on the tree T. We think of an edge as a pair of vertices, so when we use the notation $i \notin p$ for a vertex i and a matching p, we mean that i is not one of the ends of any edge in p. Let C_k denote the conjugacy class of the symmetric group S_r consisting of permutations that are the product of k disjoint transpositions, and write $\chi(C_k)$ for the common value of the character χ on such permutations. The following lemma appears in [14] and is included for completeness.

Lemma 1. *The generalized Laplacian immanantal polynomial of the tree T for the character χ is given by*

$$\sum_{k\geq 0} \chi(C_k) z^{2k} \sum_{p\in M_k(T)} \prod_{i\notin p} (x - yd_i(T))$$

where $d_i(T)$ is the degree of vertex i in the tree T.

Proof. Set $M := xI - \tilde{L}(T) = (m_{ij})$ so that the generalized Laplacian immanantal polynomial is

$$\sum_{\sigma\in S_n} \chi(\sigma) \prod_i m_{i\sigma(i)}. \tag{1}$$

The matrix entries m_{ij} are zero unless $i = j$ or there is an edge between i and j. If the permutation σ has a cycle of length 3 or greater, then corresponding term in (1) must be zero because otherwise the tree would have a loop. Therefore we need only consider permutations that are products of disjoint transpositions where, moreover, each transposition exchanges the two vertices of an edge. Such a permutation is equivalent in an obvious way to a k-matching for some k, and the lemma follows.

Lemma 2. *Two trees have the same spectrum for their generalized Laplacian if and only if they have the same generalized Laplacian immanantal polynomial for all characters.*

Proof. One direction is trivial: if two trees have the same generalized Laplacian immanantal polynomial for all characters, then their generalized Laplacians have the same characteristic polynomial and hence the same spectrum.

Conversely, if the generalized Laplacians of two trees have the same spectrum, then the characteristic polynomials of the generalized Laplacians are the same. Lemma 1 in the case χ = sgn, the fact that $\text{sgn}(C_k) = \pm 1 \neq 0$ for all k, and the fact that two equal polynomials have the same coefficients imply that the quantity

$$\sum_{p\in M_k(T)} \prod_{i\notin p} (x - yd_i(T))$$

is the same for both trees for all k. Another application of Lemma 1 completes the proof.

A sufficient condition for two trees to have the same adjacency or Laplacian spectrum

We use the phylogenetic rather than graph-theoretic definition of a subtree as follows. Define the *distal vertex* of an edge to be the vertex farthest from the root that touches that edge. Then, a subtree of a given rooted tree is what results from separating an edge from its distal vertex, which then becomes the root of the subtree.

Recall that $M_k(T)$ is the set of k-matchings of the tree T. Let $N_k(T)$ be the set of k-matchings where the chosen edges do not contact the root.

Define

$$P_k(T) := \sum_{p\in M_k(T)} \prod_{i\notin p} (x - yd_i(T))$$
$$Q_k(T) := \sum_{p\in N_k(T)} \prod_{i\notin p} (x - yd_i(T)).$$

The following lemma is implicit in [14], but again we include a proof for completeness.

Lemma 3. *Let S_1 and S_2 be trees with the same number of leaves. If $P_k(S_1) = P_k(S_2)$ and $Q_k(S_1) = Q_k(S_2)$ for all k, then any tree with S_1 as a subtree has the same generalized Laplacian spectrum as the tree obtained by substituting S_2 for S_1.*

Proof. Let T_1 be a tree with S_1 as a subtree, and write T_2 for the tree obtained by substituting S_2 for S_1. Denote by e_0 the edge that connects the rest of T_1 (resp. T_2) to the root of S_1 (resp. S_2).

We differentiate between two types of k-matchings of T_i: those that contain e_0 and those that do not. Note that a k-matching of T_i that **does not** contain e_0

restricts to an ℓ-matching of S_i for some ℓ, and all matchings of S_i arise via such a restriction. Similarly, a k-matching of T_i that does contain e_0 restricts to an ℓ-matching of S_i with the property that the root of S_i does not belong to any edge in the matching, and all matchings of S_i with this property arise via such a restriction.

Consider the formula for the characteristic polynomial of the generalized Laplacian matrix that comes from Lemma 1 with χ = sgn. Apply this formula to T_1 and T_2. The assumption $P_k(S_1) = P_k(S_2)$ (resp. $Q_k(S_1) = Q_k(S_2)$) ensures that the matchings that do not include (resp. do include) e_0 make the same contribution to the respective characteristic polynomials.

The trees depicted in Figure 1 (b) are the smallest pair of rooted bifurcating trees satisfying the criteria of this lemma. The verification of this fact was done by computer, and the corresponding P_k and Q_k polynomials are available in the code repository.

A sufficient condition for two trees to have the same distance matrix spectrum

We first recall an identity for determinants of partitioned matrices. If

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

then

$$\begin{aligned} detC &= \det\left(\begin{pmatrix} I & -C_{12}C_{22}^{-1} \\ 0 & I \end{pmatrix} C \begin{pmatrix} I & 0 \\ -C_{22}^{-1}C_{21} & I \end{pmatrix}\right) \\ &= \det\begin{pmatrix} C_{11} - C_{12}C_{22}^{-1}C_{21} & 0 \\ 0 & C_{22} \end{pmatrix} \\ &= \det(C_{22})\det(C_{11} - C_{12}C_{22}^{-1}C_{21}) \\ &= \det(C_{11})\det(C_{22} - C_{21}C_{11}^{-1}C_{21}). \end{aligned} \tag{2}$$

Lemma 4. *Form two trees T_1 and T_2 by gluing the roots of trees S_1 and S_2 with distance matrices A_1 and A_2 onto the same leaf of a common tree R. Write a_i for the column vector of distances from the leaves of S_i to the root of S_i. Suppose that the following pairs of matrices have the same spectra (where ' denotes transpose):*

$$A_i, \quad i = 1, 2,$$

$$\begin{pmatrix} A_i & a_i \\ a_i' & 0 \end{pmatrix}, \quad i = 1, 2,$$

$$\begin{pmatrix} A_i & a_i \\ 1' & 0 \end{pmatrix}, \quad i = 1, 2,$$

and

$$\begin{pmatrix} A_i & 1 \\ 1' & 0 \end{pmatrix}, \quad i = 1, 2,$$

where **1** *is a column vector with each entry* 1. *Then, the distance matrices of T_1 and T_2 have the same spectrum.*

Proof. Write B for the distance matrix of R. Then, B has the partitioned form

$$\begin{pmatrix} \breve{B} & b \\ b' & 0 \end{pmatrix},$$

where $\breve{B}$ is the distance matrix of the tree obtained from R by deleting the last leaf, b is the column vector of distances from the other leaves of R to the last leaf. Assume without loss of generality that this last leaf is the attachment point of the S_i.

Denote by D_i the distance matrix of T_i. Observe that

$$D_i = \begin{pmatrix} \breve{B} & b1' + 1a_i' \\ a_i1' + 1b' & A_i \end{pmatrix}.$$

Hence, by (2), D_i has the characteristic polynomial

$$\begin{aligned} \det(xI - D_i) &= \det(xI - A_i)\det[(xI - \breve{B}) - (-b1' - 1a_i')(xI - A_i)^{-1}(-a_i1' - 1b')] \\ &= \det(xI - A_i)\det\Big[(xI - \breve{B}) \\ &\quad - (1'(xI - A_i)^{-1}a_i)b1' \\ &\quad - (1'(xI - A_i)^{-1}1)bb' \\ &\quad - (a_i'(xI - A_i)^{-1}a_i)11' \\ &\quad - (a_i'(xI - A_i)^{-1}1)1b'\Big]. \end{aligned}$$

Using (2) again, we see that a partitioned matrix of the form

$$\begin{pmatrix} A & g \\ h' & 0 \end{pmatrix},$$

where g and h are column vectors, has characteristic polynomial

$$\det(xI - A)\left[x - h'(xI - A)^{-1}g\right],$$

and the result follows.

It was verified by computer that the trees in Figure 1 (b) are the smallest such that have distance matrices A_i and vectors a_i satisfying the criteria of this lemma. The corresponding characteristic polynomials are available in the code repository. We note with surprise that the smallest pair of trees with the exchange property for the distance matrix are the same as the smallest pair with the exchange property for the generalized Laplacian; this is a curiosity for which we do not have an explanation.

Asymptotic numbers of trees

As outlined in the Introduction, the proof of Theorem 1 follows immediately from Lemma 2, Lemma 3, Lemma

4, the existence of the trees in Figure 1 (b), and the following result.

Proposition 1. *Let T be a rooted tree. Let t_n be the number of trees with n leaves. Let s_n be the number of such trees that do not contain T as a rooted subtree. Then, the fraction s_n/t_n goes to zero as n goes to infinity.*

Proof. Suppose that T has a leaves. Let $f(x) := \sum_{i=1}^{\infty} t_i x^i$ and $f_a(x) := \sum_{i=1}^{\infty} s_i x^i$ denote the generating functions for t_n and s_n, respectively. Write ρ for the radius of convergence of the power series f and ρ_a for the radius of convergence of the power series f_a. Note that $\rho \le \rho_a < 1$.

It is shown in [26] that $\rho = 0.402698\ldots$ and

$$\lim_{n\to\infty} n^{3/2}\rho^n t_n = \eta,$$

where $\eta = 0.7916032\ldots$ (see [27] for an asymptotic expansion of t_n that extends this result and [28-31] for reviews of general methods for determining asymptotic numbers of trees of various sorts from a knowledge of the functional equations that their generating functions solve). Since s_n is $o(\alpha^{-n})$ for any $0 < \alpha < \rho_a$, it follows that s_n/t_n is $o(\beta^n)$ for any $\beta > \rho/\rho_a$, and the proposition will hold if we can show that $\rho < \rho_a$.

For the sake of completeness and because it serves as a good introduction to the derivation of the functional equation satisfied by the generating function of s_n, we first derive the well-known functional equation satisfied by the generating function of the t_n. See the comments after the proof of the lemma for some remarks about the history of the latter generating function.

By decomposing a tree into the two subtrees rooted at the daughters of the root, it is clear that

$$t_n = t_1 t_{n-1} + t_2 t_{n-2} + \cdots + t_{m-1} t_{m+1}, \quad \text{for } n = 2m+1,$$
$$t_n = t_1 t_{n-1} + t_2 t_{n-2} + \cdots + t_{m-1} t_{m+1} + \binom{t_m}{2} + t_m, \quad \text{for } n = 2m.$$

These expressions are equivalent to the statement

$$\sum_{i=1}^{n-1} t_i t_{n-i} = 2t_n - t_{n/2} \tag{3}$$

where $t_{n/2}$ is set to zero if n is odd.

From (3) the generating function f satisfies the functional equation

$$\begin{aligned} f^2(x) &= \sum_{n=2}^{\infty} x^n \sum_{i=1}^{n-1} t_i t_{n-i} \\ &= \sum_{n=2}^{\infty} x^n (2t_n - t_{n/2}) \\ &= 2f(x) - f(x^2) - 2x. \end{aligned}$$

It will be convenient to consider the function $g := 1 - f$, which satisfies the functional equation

$$g(x^2) = 2x + g^2(x). \tag{4}$$

It is shown in [15] that:

- The radius of convergence ρ is strictly positive.
- The functional equation (4) has a unique solution in the whole complex plane, and this solution agrees with our power series in $\{x \in \mathbb{C}: |x| < \rho\}$.
- If, with a slight abuse of notation, we also denote this solution by g, then $g(\rho) = 0$.
- The point ρ is the only zero of g within $\{x \in \mathbb{C} : |x| < 1\}$.

It is clear from the power series that g is continuous and decreasing on $[0, \rho)$ and $g(0) = 1$. Hence g is strictly positive on $[0, \rho)$.

As observed in [15], these observations suggest a method for computing ρ. Put $h(x) = g(x)/\sqrt{x}$, so that h satisfies $h(x^2) = 2 + h^2(x)$. Set

$$w_k(\eta) := \left(2 + \eta^{2^k}\right)^{2^{-k}}, \quad \eta \in \mathbb{R},$$

and

$$q_n := w_{n-1} \circ w_{n-2} \cdots \circ w_0,$$

so that each function q_n is strictly increasing on $[-2, +\infty)$ and $q_1 \le q_2 \le \ldots$. In particular, $\lim_{n\to\infty} q_n(y)$ exists in $\mathbb{R} \cup \{+\infty\}$ for each $y \in [-2, +\infty)$. Moreover,

$$\lim_{n\to\infty} q_n(h^2(x)) = \lim_{n\to\infty} (h(x^{2^n}))^{2^{1-n}} = \lim_{n\to\infty} \frac{(g(x^{2^n}))^{2^{1-n}}}{x} = \frac{1}{x}$$

for $0 < x < 1$. Therefore

$$\frac{1}{\rho} = \lim_{n\to\infty} q_n(0).$$

Conveniently, (3) holds with s_n in place of t_n for all n except for $n = a$; in this case one simply adds two to the right hand side of the equation to make up for the fact that $s_a = t_a - 1$. Hence f_a satisfies the functional equation.

$$f_a^2(x) = 2f_a(x) - f_a(x^2) - 2x + 2x^a. \tag{5}$$

Set $g_a := 1 - f_a$, so that

$$g_a(x^2) = 2x - 2x^a + g_a^2(x). \tag{6}$$

It is clear that g_a is continuous and decreasing on $[0, \rho_a)$ and $g_a(0) = 1$. Following the arguments in [15], the functional equation (6) has a unique solution in the

whole complex plane, and this solution agrees with our power series in $\{x \in \mathbb{C}: |x| < \rho_a\}$. Moreover, analogues of the other properties of g obtained in [15] hold for g_a.

Set $h_a(x) = g_a(x)/\sqrt{x}$, so that

$$h_a(x^2) = 2 - 2x^{a-1} + h_a^2(x).$$

Put

$$w_{k,a,\xi}(\eta) := \left(2 - 2\xi^{2^k} + \eta^{2^k}\right)^{2^{-k}}, \quad \eta \in \mathbb{R},$$

and

$$q_{n,a,\xi} := w_{n-1,a,\xi} \circ w_{n-2,a,\xi} \cdots \circ w_{0,a,\xi}.$$

Then

$$\lim_{n\to\infty} q_{n,a,x^{a-1}}(h_a^2(x)) = \lim_{n\to\infty} (h_a(x^{2^n}))^{2^{1-n}} = \lim_{n\to\infty} \frac{(g_a(x^{2^n}))^{2^{1-n}}}{x} = \frac{1}{x}$$

for $0 < x < 1$, and, in particular,

$$\frac{1}{\rho_a} = \lim_{n\to\infty} q_{n,a,\rho_a^{a-1}}(0).$$

Now

$$\begin{aligned} q_{n,a,\rho_a^{a-1}}(0) &= w_{n-1,a,\rho_a^{a-1}} \circ w_{n-2,a,\rho_a^{a-1}} \cdots \circ w_{0,a,\rho_a^{a-1}}(0) \\ &\le w_{n-1} \circ w_{n-2} \cdots \circ w_1 \circ w_{0,a,\rho_a^{a-1}}(0) \\ &= q_n\left(-2\rho_a^{a-1}\right), \end{aligned}$$

and so

$$\frac{1}{\rho_a} \le \lim_{n\to\infty} q_n(-2\rho_a^{a-1}) \le \lim_{n\to\infty} q_n(0) = \frac{1}{\rho}.$$

It therefore suffices to show that the function $y \mapsto \lim_{n\to\infty} q_n(y)$ is strictly increasing on $(-\varepsilon, +\infty)$ for some $0 < \varepsilon < 2$.

Observe that the derivative of q_n satisfies

$$q_n' = \prod_{k=1}^{n-1} w_k' \circ q_k.$$

For $k \ge 1$,

$$\begin{aligned} w_k'(x) &= x^{2^k-1}\left(2 + x^{2^k}\right)^{2^{-k}-1} \\ &= \left(2x^{-2^k} + 1\right)^{2^{-k}-1}, \end{aligned}$$

so that $x \mapsto w_k'(x)$ is non-decreasing for $x > 0$. For $y \in (-\varepsilon, +\infty)$,

$$w_k' \circ q_k(y) \ge w_k' \circ q_1(y) \ge w_k' \circ q_1(-\varepsilon) = w_k'(2 - \varepsilon)$$

and hence

$$\liminf_{n\to\infty} \inf_{y>-\varepsilon} q_n'(y) \ge \prod_{k=1}^{\infty} w_k'(2 - \varepsilon).$$

Taking $0 < \varepsilon < 1$, the proof will be completed by demonstrating for any $x > 1$ that

$$\prod_{k=1}^{\infty} w_k'(x) > 0.$$

Taking the logarithm gives

$$\begin{aligned} \sum_{k=1}^{\infty} (2^{-k} - 1)\log\left(2x^{-2^k} + 1\right) &> -\sum_{k=1}^{\infty} \log\left(2x^{-2^k} + 1\right), \\ &> -\sum_{k=1}^{\infty} 2x^{-2^k}, \end{aligned}$$

and this series clearly converges by the ratio test.

In relation to Proposition 1, we note from [13] that the number of rooted strictly bifurcating trees without a given subtree is asymptotically smaller than the number of all graph-theoretic trees (see also [32] for more about the enumeration of general trees without a given subtree), but this is not enough for our purposes. We needed to show that it is asymptotically smaller than the space of all rooted strictly bifurcating trees. The generating function for t_n seems first to have been investigated in [15] in connection with enumerating "types of arrangements" in a commutative but non-associative algebra, such as $a_1(a_2(a_3a_4))$ or $(a_1a_2)(a_3a_4)$; these are identical to rooted bifurcating trees in the "Newick" format [1]. The recursion behind the generating function has been re-discovered independently several times such as in [33] - see [34] for a discussion and several further references. We remark that numerically iterating the quantity q_n of the proof converges quickly to the value of ρ^{-1} calculable by other means [26,35]. We also observe that the methods of [27-31] can be used to show, in the notation of the proof, that $\lim_{n\to\infty} n^{3/2}\rho_a^n s_n = \eta_a$ for some positive constant η_a and hence $\lim_{n\to\infty}(\rho_a/\rho)^n(s_n/t_n) = \eta_a/\eta$, but we don't pursue this matter here.

Numerical experiments

Proposition 1 says nothing about the rate of convergence of the fraction. Here we investigate this rate using computation. The characteristic polynomials for the generalized Laplacian were calculated via a doubly-recursive algorithm to enumerate matchings. The characteristic polynomials for the distance matrices were calculated via the Leverrier-Faddeev algorithm [36].

Table 1 shows that the fraction of trees with unique spectra does not go to zero very quickly. We can't compute this fraction for large numbers of leaves, but we can get some idea of the convergence by using the recursion relation corresponding to the generating function (5). Figure 2 shows the number of trees that do not have one of two subtrees of size seventeen as a subtree. This is an actual fraction that can be used with Proposition 1 in order to prove Theorem 1 for the generalized Laplacian matrix.

Figure 2 shows that this fraction converges extremely slowly, despite the fact that as shown above it is asymptotically equivalent to β^n for some $0 < \beta < 1$. It is important to note, however, that this fraction is probably a very crude upper bound on the fraction of trees that share a spectrum with another tree. As can be seen in Table 1, the actual number not sharing a spectrum goes down considerably more quickly, though it is probably still the case that the vast majority of trees of intermediate size should have their own spectra.

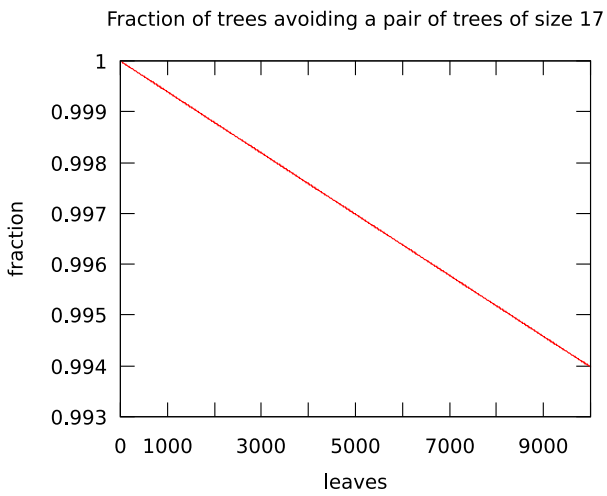

Figure 2 The fraction of trees not containing a given pair of subtrees of size 17 .

Conclusions

Spectral invariants of matrix formulations of trees are a natural way to quantify the shape of phylogenetic trees. However, in this paper we show that a complete classification of tree shapes using common spectral invariants of generalized Laplacian and distance matrices is not possible. For either of these choices of matrix we show that the fraction of binary trees with a unique spectrum goes to zero as the number of leaves goes to infinity, but the rate of convergence of the above fraction to zero appears to be slow. For the adjacency and Laplacian matrices, we show that the *a priori* more informative immanantal polynomials have no greater power to distinguish between trees.

Table 1 The number of trees, the number of spectra for the generalized Laplacian (GLS), and the number of spectra for the distance matrix (DS).

leaves	trees	GLS	DS
2	1	1	1
3	1	1	1
4	2	2	2
5	3	3	3
6	6	6	6
7	11	11	11
8	23	22	23
9	46	45	46
10	98	95	98
11	207	203	207
12	451	443	451
13	983	972	983
14	2179	2159	2179
15	4850	4827	4850
16	10905	10870	10905
17	24631	24580	24630
18	56011	55931	56009
19	127912	127830	127908

Acknowledgements

SNE is supported in part by NSF grants DMS-0405778 and DMS-0907630, and part of the research was conducted during a visit to the Pacific Institute for the Mathematical Sciences.

Author details

[1]Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [2]Department of Statistics, University of California at Berkeley, Berkeley, California, USA.

Authors' contributions

FAM conceived of the project, proved the theorems, performed the numerical experiments, and wrote the paper. SNE applied the immanant, proved the theorems, and wrote the paper.

Competing interests

The authors declare that they have no competing interests.

Received: 5 October 2011 Accepted: 21 May 2012
Published: 21 May 2012

References

1. Felsenstein J: *Inferring Phylogenies* Sunderland, MA: Sinauer Press; 2004.
2. Mooers A, Heard S: **Evolutionary process from phylogenetic tree shape.** *Q Rev Biol* 1997, **72**:31-54.
3. Kirkpatrick M, Slatkin M: **Searching for evolutionary patterns in the shape of a phylogenetic tree.** *Evolution* 1993, **47(4)**:1171-1181.
4. Agapow P, Purvis A: **Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis.** *Syst Biol* 2002, **51(6)**:866-872.
5. Matsen FA: **A geometric approach to tree shape statistics.** *Systematic biology* 2006, **55(4)**:652-661.
6. Matsen FA: **Optimization over a class of tree shape statistics.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2007, **4(3)**:506-512.

7. Biggs N: *Algebraic graph theory*. second edition. Cambridge Mathematical Library, Cambridge: Cambridge University Press; 1993.
8. Chung FRK: *Spectral graph theory, Volume 92 of CBMS Regional Conference Series in Mathematics* Published for the Conference Board of the Mathematical Sciences, Washington, DC; 1997.
9. Semple C, Steel M: *Phylogenetics, Volume 24 of Oxford Lecture Series in Mathematics and its Applications* Oxford: Oxford University Press; 2003.
10. Gupta A: **Embedding tree metrics into low-dimensional Euclidean spaces.** *Discrete Comput Geom* 2000, **24**:105-116.
11. Matoušek J: *Lectures in Discrete Geometry* New York: Springer; 2002.
12. Cvetković DM, Doob M, Sachs H: *Spectra of graphs*. third edition. Heidelberg: Johann Ambrosius Barth; 1995.
13. Schwenk AJ: **Almost all trees are cospectral.** *New Directions in the Theory of Graphs* New York: Acade-meic Press; 1973, 275-307.
14. Botti P, Merris R: **Almost all trees share a complete set of immanantal polynomials.** *J Graph Theory* 1993, **17(4)**:467-476.
15. Wedderburn JHM: **The functional equation $g(x^2) = 2ax + [g(x)]^2$.** *Ann of Math (2)* 1922, **24(2)**:121-140.
16. Robinson GdB: *Representation theory of the symmetric group* University of Toronto Press, Toronto; 1961, Mathematical Expositions, No. 12.
17. Fulton W, Harris J: *Representation theory, Volume 129 of Graduate Texts in Mathematics* New York: Springer-Verlag; 1991.
18. Simon B: *Representations of finite and compact groups, Volume 10 of Graduate Studies in Mathematics* Providence, RI: American Mathematical Society; 1996.
19. Sagan BE: *The symmetric group, Volume 203 of Graduate Texts in Mathematics*. second edition. New York: Springer-Verlag; 2001.
20. Littlewood DE, Richardson AR: **Group characters and algebra.** *Philos Trans Roy Soc London A* 1934, **233**:99-141.
21. Littlewood DE: *The Theory of Group Characters and Matrix Representations of Groups* New York: Oxford University Press; 1940.
22. Steyaert JM, Flajolet P: **Patterns and pattern-matching in trees: an analysis.** *Inform and Control* 1983, **58(1-3)**:19-58.
23. Graham R, Lovász L: **Distance matrix polynomials of trees.** *Adv Mathematics* 1978, **29**:60-88.
24. Bhamidi S, Evans SN, Sen A: **Spectra of large random trees.** *U.C Berkeley Department of Statistics Technical Report No. 771* 2009, [To appear in J. Theoret. Probab.].
25. Chailloux E, Manoury P, Pagano B: **Développement d'applications avec Objective CAML.** Sebastopol, CA: O'Reilly 2000;[http://caml.inria.fr/pub/docs/oreilly-book/], .
26. Otter R: **The number of trees.** *Ann of Math (2)* 1948, **49**:583-599.
27. Landau BV: **An asymptotic expansion for the Wedderburn-Etherington sequence.** *Mathematika* 1977, **24(2)**:262-265.
28. Harary F, Robinson RW, Schwenk AJ: **Twenty-step algorithm for determining the asymptotic number of trees of various species.** *J Austral Math Soc Ser A* 1975, **20(4)**:483-503.
29. Harary F, Robinson RW, Schwenk AJ: **Corrigendum: "Twenty-step algorithm for determining the asymptotic number of trees of various species" [J. Austral. Math. Soc. Ser. A 20 (1975), no. 4, 483-503; MR0406858 (53 #10644)].** *J Austral Math Soc Ser A* 1986, **41(3)**:325.
30. Bender EA: **Asymptotic methods in enumeration.** *SIAM Rev* 1974, **16**:485-515.
31. Bender EA: **Errata: "Asymptotic methods in enumeration" (SIAM Rev. 16 (1974), 485-515).** *SIAM Rev* 1976, **18(2)**:292.
32. Lu T: **The enumeration of trees with and without given limbs.** *Discrete Math* 1996, **154(1-3)**:153-165.
33. Etherington I: **Non-associate powers and a functional equation.** *Math Gaz* 1937, **21**:36-39.
34. Olds CD, Becker HW: **Advanced Problems and Solutions: Solutions: 4277.** *Amer Math Monthly* 1949, **56(10)**:697-699.
35. Harding EF: **The probabilities of rooted tree-shapes generated by random bifurcation.** *Adv Appl Probability* 1971, **3**:44-77.
36. Hou SH: **A simple proof of the Leverrier-Faddeev characteristic polynomial algorithm.** *SIAM Rev* 1998, **40(3)**:706-709.

doi:10.1186/1748-7188-7-14
Cite this article as: Matsen and Evans: Ubiquity of synonymity: almost all large binary trees are not uniquely identified by their spectra or their immanantal polynomials. *Algorithms for Molecular Biology* 2012 **7**:14.