Phylogenetic Invariants for the Uniform Evolution Model

Radu H. Mihaescu

12/08/2003

1 The Model of Uniform Evolution

In this project we will look at the properties of phylogenetic trees when the model of evolution we assume is uniform, namely the transition probability matrix is the same at all edges of the tree. We will see that in this case a number of linear algebraic invariants arise naturally from the model and can even fully describe the topology of the tree when all the leafs of the tree are situated at the same depth. Finally we will present the results of some numerical experiments that were conducted using the software package Macaulay I.

Suppose that we have a phylogenetic tree \( T \) with root \( R \) and that space of character states is \( S \), with \(|S| = k\). Throughout the rest of this paper we will assume that the transition probability matrix associated with every edge is the same, namely the \( k \times k \) matrix \( A \). Let \( v \) be a vertex of \( T \) and let us denote by \( P_v \) probability distribution of character states at \( v \), where by \( \chi(v) \) we denote the character state at \( v \). Let \( d(v) \) be the depth of \( v \) in the tree \( T \) (\( d(v) \) is the length of the unique path from \( v \) to the root \( R \)). Then we have the following relation:

\[
P_v = P_R \cdot A^{d(v)}
\]

Now let us make the additional assumption that the matrix \( A \) is ergodic, meaning that all its entries strictly larger than some positive \( \epsilon \). We obtain the following well-known inequality:

\[
(1 - \epsilon) \cdot d(v, w) \geq d(vA, wA),
\]

where \( d(v, w) \) is the distance between the two \( k \)-dimensional vectors \( v \) and \( w \), under the \( L_\infty \) norm. We can see that any sequence of probability distributions \( p_n = p_0 \cdot A^n \) will therefore form a Cauchy sequence, and its limit \( \overline{p} \) will be a stable distribution for \( A \), namely \( \overline{p}A = \overline{p} \). Also, \( \overline{p} \) is the only stable distribution for \( A \), which is easy to infer from the above inequality.

We can now see that if \( P_R \neq \overline{p} \), then perfect information (or infinite data), would allow us to distinguish if \( d(v) = d(w) \) by simply observing that \( P_v = P_w \). On the other hand, when trees become very deep, the probability distributions associated with their leafs will all be close to \( \overline{p} \). Therefore, under imperfect data, we will not be able to make use of this observation, unless the errors in the data are small enough to compensate for the depth of the tree.

Given a tree \( T \), an algebraic invariant of \( T \) is a polynomial in the entries of the joint probability distribution of character states at the leafs of \( T \), which evaluates at zero for all probability distributions generated by \( T \), regardless of the model parameters, namely the specific value of the transition matrix \( A \) and the root distribution \( P_R \). One such example is given by the trivial observation that all joint probabilities naturally sum up to one. This is called the trivial invariant. In general, for a specific model of evolution, we will call invariants that hold for all trees, like the one given above,
model invariants, whereas invariants that hold only for a specific subset of the set of possible trees are called phylogenetically informative invariants.

2 Linear Invariants

First of all, we will introduce some more notation. First denote by $\chi(X)$ the set of character states at the vertices in the set $X$. For two nodes $a,b$ in the tree $T$, let $R_{a,b}$ be their nearest common ancestor and let $T_{a,b}$ be the subtree of $T$ that is rooted at $R_{a,b}$. Also let $L(T)$ be the set of leafs of $T$ and let $L_{a,b} = L(T) - L(T_{a,b})$. Now suppose $a$ and $b$ are two leafs of the tree $T$ such that $d(a) = d(b)$. Then

$$p(\chi(a) = \alpha, \chi(b) = \beta | L_{a,b}) = p(\chi(a) = \beta, \chi(b) = \alpha | L_{a,b}),$$

so

$$\sum_{\chi(L(T_{a,b}) - \{a,b\})} p(\chi(L(T_{a,b}) - \{a,b\}), \chi(a) = \alpha, \chi(b) = \beta, \bar{\chi}(L_{a,b})) =$$

$$\sum_{\chi(L(T_{a,b}) - \{a,b\})} p(\chi(L(T_{a,b}) - \{a,b\}), \chi(a) = \beta, \chi(b) = \alpha, \bar{\chi}(L_{a,b}))$$

(2)

In other words, marginalizing over the other leafs of the tree $T_{a,b}$ allows us to treat the pair $a,b$ as a cherry of a tree minor of $T$. Since $d(a) = d(b)$, the transition probability matrix from $R_{a,b}$ to $a$ is the same as that from $R_{a,b}$ to $b$, and therefore $\chi(a)$ and $\chi(b)$ can be interchanged.

To exemplify, let us consider the following trees

Let $p_{xyz} = p(\chi(l_1) = x, \chi(l_2) = y, \chi(l_3) = z)$. Then, for the both of the above trees, $p_{xij} = p_{xji}$, but $p_{xij} \neq p_{ixj}$. Note that we need not require that the tree $T$ be binary. The same argument works also for non-binary trees. For instance, for the tree in the figure below we have $p_{xijk} = p_{x\sigma(i)\sigma(j)\sigma(k)}$, for any permutation $\sigma$ of $i,j,k$.
So when are the invariants described by equation (2) enough for us to deduce the entire topology of $T$? Each such invariant tells us that two vertices $a$ and $b$ are at the same depth, and it also specifies which leafs of $T$ are in $T_{a,b}$ and which leafs are not, since we are marginalizing over leafs in $T_{a,b}$, while holding the characters at the other leafs fixed. We can therefore see that in the case when the subtrees of $T$ rooted at the children of the root $R(T)$ are perfectly balanced (all leafs are at the same depth), the linear invariants of the type given in equation (2) fully specify the topology of the entire tree $T$.

This condition can sometimes arise very naturally. For instance, the case of perfectly balanced trees corresponds to the assumption of a molecular clock. If we assume that each species evolves at the same rate, all present day species will be at the same distance from their common ancestor, and therefore at the same depth in the phylogenetic tree.

3 Experimental Results

Given the analysis of the previous sections, several questions arise naturally. For instance, are the invariants described by equation (2) all the linear invariants that can occur on a tree $T$? Can we pursue the same approach to obtain invariants of higher degree? How is the number of invariants of a certain degree related to the complexity of the tree?

In order to gain more intuition about these questions, we have conducted a series of experiments on a set of eleven small trees (up to ten edges). This was done using the computational algebra package Macaulay I. The model we assume is very simple, as the computations become intractable very soon for more complicated ones. Specifically, we assume that there are only two possible character states, 0 and 1, and that the root $R$ of the tree $T$ has uniform distributions of character states: $P_R = (0.5, 0.5)$.

The table on the following page summarizes our results. The question marks on the last two rows signify the fact that finding all the invariants for the last two trees proved to be computationally very intensive, and the programs were stopped before producing the entire calculation.

As we can see from the table, the two HMM-like trees at the bottom at the table do have additional linear invariants that do not arise from those specified by equation (2). Indeed, for the 10th tree, one would expect only eight linear invariants, given by switching the two states of the bottom cherry nodes, while holding the states of the other three fixed. However, we obtain two additional such linear invariants, namely:

\[ I_1 = p_{01001} - p_{01011} - p_{01100} + p_{01110} - p_{10010} + p_{10011} + p_{10100} - p_{10110} \]

\[ I_2 = p_{00010} + p_{00011} - p_{00100} - p_{00111} + 2p_{01011} - 2p_{01110} - p_{01100} + 2p_{10100} + 2p_{11000} + p_{11011} - p_{11010} - p_{11110} \]

In the case of the HMM tree on 6 leafs (tree no. 12), we also obtain thirty one linear invariants, instead of just the sixteen generated by equation 2. Out of the trees we have studied, these two HMM-like trees are in fact the only ones where additional linear invariants arise. This may in fact be caused by our assumption of an uniform distribution of character states at the root. Results of more simulations where this condition is relaxed will follow in a forthcoming report.

One additional observation is that in the cases where the trees are fairly balanced, there are very few invariants of degree 2, whereas in the case of highly unbalanced trees like no. 4, 10 and 11 we observe an abundance of degree 2 invariants. This seems to suggest that the low degree invariants
should contain much information about the structure of the trees, with the invariants of degree two compensating for the cases when the linear ones are lacking.

Finally, we observe that the base locus of the ideal of joint probability distributions has always codimension 3, while its degree seems to be growing approximately linearly with the size of the tree.

<table>
<thead>
<tr>
<th>TREE</th>
<th>No. of invariants of deg:</th>
<th>Base Locus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 +</td>
<td>codim</td>
</tr>
<tr>
<td>1</td>
<td>2 0 5 9 3 1 0 0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3 0 0 0 5 11 2 0</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>10 0 1 7 6 1 0 0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4 27 34 3 0 0 0 0</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>6 2 66 25 4 0 0 0</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>5 13 50 9 1 0 0 0</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>6 0 72 37 6 1 0 0</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>20 10 92 18 6 0 0 0</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>20 1 155 28 2 1 0 0</td>
<td>3</td>
</tr>
</tbody>
</table>