MAXIMUM LIKELIHOOD SUPERTREES

MIKE STEEL AND ALLEN RODRIGO

ABSTRACT. We analyse a maximum-likelihood approach for combining phylo-genetic trees into a larger 'supertree'. This is based on a simple exponential model of phylogenetic error, which ensures that ML supertrees have a simple combinatorial description (as a median tree, minimising a weighted sum of distances to the input trees). We show that this approach to ML supertree reconstruction is statistically consistent (it converges on the true species su-pertree as more input trees are combined), in contrast to the widely-used MRP method, which we show can be statistically inconsistent under the exponential error model. We also show that this statistical consistency extends to an ML approach for constructing species supertrees from gene trees. In this setting, incomplete lineage sorting (due to coalescence rates of homologous genes be-ing lower than speciation rates) has been shown to lead to gene trees that are frequently different from species trees, and this can confound efforts to reconstruct the species phylogeny correctly.

1. INTRODUCTION

Combining trees on different, overlapping sets of taxa into a parent 'supertree' is now a mainstream strategy for constructing large phylogenetic trees. The literature on supertrees is growing steadily: new methods of supertree reconstruction are be-ing developed (Cotton and Wilkinson, 2007) and supertree analyses are shedding light on fundamental evolutionary questions (Bininda-Emonds et al., 2007). De-spite this surge in research activity, it is probably fair to say that biologists are still confused about what supertrees really are and what it is we do when we build a supertree. Are we, as some maintain, simply summarising the phylogenetic in-formation contained in a group of subtrees? Or are we trying to derive the best estimate of phylogeny given the information at hand? Nor is it clear which of these two conceptually different objectives underpin the various supertree reconstruction methods.

We take the view that what biologists really want a supertree reconstruction method to deliver is the best hypothesis of evolutionary relationships that can be inferred from the data available. Obviously, it is not the case that the supertree con-structed as a summary statistic will necessarily be the best estimate of phylogeny.

Nonetheless, if we are prepared to consider supertree reconstruction a problem of phylogenetic *estimation*, we have at our disposal an arsenal of phylogenetic tools and methods that have been tried and tested. Matrix Representation with Parsimony (MRP; (Baum and Ragan, 1992)), Matrix Representation with Compatibility (MRC; (Rodrigo, 1996; Ross and Rodrigo, 2004)) and, most recently, Bayesian supertree reconstruction (BSR, (Ronquist et al., 2004)) are undoubtedly inspired by standard phylogenetic methods. A gap remains, though, as there has been remarkably little development of likelihood-based methods for supertree reconstruction.

In this paper, we analyse one approach to obtain maximum-likelihood (ML) estimates of supertrees, based on a probability model that permits 'errors' in subtree topologies. The approach is of the type described by Cotton and Page (2004), and it permits supertrees to be estimated even if there is topological conflict amongst the constituent subtrees. We show that ML estimates of supertrees so obtained are statistically consistent under fairly general conditions. By contrast, we show that MRP may be inconsistent under these same conditions. We then consider a further complication that arises in the supertree setting when combining gene trees into species trees - in addition to the possibility that the input gene trees are reconstructed incorrectly (either a consequence of the reconstruction method used, or some sampling error), there is a further stochastic process that leads to the (true) gene trees differing from their underlying species tree (a consequence of incomplete lineage sorting). Although simple majority-rule approaches (and gene concatenation) have recently been shown to be misleading, we show that an ML supertree approach for combining gene trees is also statistically consistent.

1.1. **Terminology.** Throughout this paper, unless stated otherwise, phylogenetic trees may be either rooted or unrooted, and we will mostly follow the notation of Semple and Steel (2003). In particular, given a (rooted or unrooted) phylogenetic tree $\mathcal{T}$ on a set X of taxa (which will always label the leaves of the tree), any subset Y of X induces a phylogenetic tree on taxon set Y, denoted $\mathcal{T}|Y$, which, informally, is the subtree of $\mathcal{T}$ that connects the taxa in Y only. In the *supertree* problem, we have a sequence $\mathcal{P} = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k)$ of input trees, called a *profile*, where $\mathcal{T}_i$ is a phylogenetic tree on taxon set X_i. We wish to combine these trees into a phylogenetic tree $\mathcal{T}$ on the union of the taxon sets (i.e. $X = X_1 \cup X_2 \cup \cdots \cup X_k$). We assume that the trees in $\mathcal{P}$ are either all rooted or all unrooted, and that $\mathcal{T}$ is rooted or unrooted accordingly. We will mostly assume that trees are *fully-resolved* (i.e. binary trees, without polytomies); in Section 5 we briefly describe how this restriction can be lifted.

A special case of the supertree problem arises when the taxon sets of the input trees are all the same ($X_1 = X_2 = \cdots = X_k$). This is the much studied *consensus tree* problem. In an early paper McMorris (1990) described how, in this consensus setting, the majority rule consensus tree can be given a maximum likelihood interpretation. However this approach is quite different to the one described here (even when restricted to the consensus problem).

In this paper, we will denote the underlying ('true') species tree as $\mathcal{T}_0$ (assuming that such a tree exists and that the evolution of the taxa has not involved reticulate

processes such as the formation of hybrid taxa). In an ideal world, we would like $\mathcal{T}_i = \mathcal{T}_0|X_i$ for each tree $\mathcal{T}_i$ in the profile – that is, we would like each of the reconstructed trees to be identical to the subtree of the 'true' tree for the taxa in X_0. But in practice, the trees $\mathcal{T}_1, ..., \mathcal{T}_k$ are unlikely to even be compatible (i.e. no phylogenetic tree $\mathcal{T}$ exists for which $\mathcal{T}_i = \mathcal{T}|X_i$ for all i).

2. An exponential model of phylogenetic error

Species trees that have been inferred from data may differ from the true underlying species tree for numerous reasons, including sampling effects (short and/or site-saturated sequences, or poorly defined characters), model violation, sequencing or alignment errors, and so forth. In this section, we will assume a simple model of phylogenetic error in which the probability of observing a given tree falls off exponentially with its 'distance' from an underlying generating tree (e.g. the true species tree $\mathcal{T}_0$). This type of model has been described by Holmes (2003), albeit from a different perspective. Suppose d is some metric on resolved phylogenetic trees. In the exponential model, the probability, denoted $\mathbb{P}_{\mathcal{T}}[\mathcal{T}']$, of reconstructing any species tree $\mathcal{T}'$ on taxon set Y, when $\mathcal{T}$ is the generating tree (on taxon set X) is proportional to an exponentially decaying function of the distance from $\mathcal{T}'$ to $\mathcal{T}|Y$. In other words,

$$\mathbb{P}_{\mathcal{T}}[\mathcal{T}'] = \alpha \exp(-\beta d(\mathcal{T}', \mathcal{T}|Y)). \tag{1}$$

The constant β can vary with Y and other factors (such as the quality of the data); for example, trees constructed from long high-fidelity sequences are likely to have a larger β than trees constructed from shorter and/or noisier sequences. The constant α is simply a normalising constant to ensure that $\sum_{\mathcal{T}'} \mathbb{P}_{\mathcal{T}}[\mathcal{T}'] = 1$, where the sum is over all fully resolved phylogenetic trees $\mathcal{T}'$ on taxon set Y. When we have a sequence $(X_1, X_2, \ldots)$ of subsets of X, we will reflect the dependence of α, β on X_i by writing α_i, and β_i. Note that α_i is determined entirely by β_i and $|X_i|$.

Note that, implicit in (1), the probability of $\mathcal{T}'$ depends only on the subtree of $\mathcal{T}$ connecting the species in $\mathcal{T}'$ and not on the other species in $\mathcal{T}$ that are not present in $\mathcal{T}'$.

Now, suppose we observe the profile of trees $\mathcal{P} = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k)$ as above, where $\mathcal{T}_i$ has leaf set X_i. Assume that, for each i, the tree $\mathcal{T}_i$ has been independently sampled from the exponential distribution (1) with $\beta = \beta_i$. Select a phylogenetic X–tree $\mathcal{T}$ that maximises the probability of generating the observed profile $\mathcal{P}$ (we call this type of a tree $\mathcal{T}$ an *ML supertree for* $\mathcal{P}$). In the special case where d is the nearest-neighbor interchange (NNI) metric, and the β_i values are all equal, this ML supertree was described by (Cotton and Page, 2004). An ML supertree has a simple combinatorial description as a (weighted) median tree, as the following result shows.

Proposition 2.1. *For any metric d on phylogenetic trees, an ML supertree for a profile $\mathcal{P}$ is precisely a tree $\mathcal{T}$ that minimises the weighted sum:*

$$\sum_{i=1}^{k} \beta_i d(\mathcal{T}_i, \mathcal{T}|X_i).$$

Proof. By the independence assumption

$$\mathbb{P}_\mathcal{T}[(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k)] = \prod_{i=1}^{k} \mathbb{P}_\mathcal{T}[\mathcal{T}_i],$$

and, by (1), $\mathbb{P}_\mathcal{T}[\mathcal{T}_i] = \alpha_i \exp(-\beta_i d(\mathcal{T}_i, \mathcal{T}|X_i))$. Consequently, $\mathbb{P}_\mathcal{T}[(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k)]$ is proportional to

$$\exp(-\sum_{i=1}^{k} \beta_i d(\mathcal{T}_i, \mathcal{T}|X_i)),$$

and this is maximised for any tree $\mathcal{T}$ that minimises $\sum_{i=1}^{k} \beta_i d(\mathcal{T}_i, \mathcal{T}|X_i)$. This completes the proof. □

Notice that in the consensus tree setting, and where d is the symmetric difference (Robinson-Foulds) metric, the consensus of the ML supertrees is the same as the usual majority rule consensus tree. This follows from earlier results by Bathélemy and McMorris (1986) (see also Cotton and Wilkinson, 2007).

3. Statistical Consistency of ML supertrees under the exponential model

Is the ML procedure statistically consistent as the number (k) of trees in the profile grows? More precisely, under what conditions is the method guaranteed to converge on the underlying generating tree $\mathcal{T}_0$ as we add more trees to the analysis? The problem is slightly different from other settings (such as the consistency of ML for tree reconstruction from aligned sequence data) where one has a sequence of i.i.d. random variables. In the supertree setting, it is perhaps unrealistic to expect that the data-sets are generated according to an identical process, since the sequence of subsets $X_1, X_2, \ldots$ of X is generally deliberately selected.

To formalise the statistical consistency question in this setting, let $X_1, X_2, \ldots,$ be a sequence of subsets of X. It is clear that the X_i's must cover X in some 'reasonable' way in order for the ML supertree method to be consistent – for example, if some taxon is not present in any X_i, or is present in only a small number of input trees, then we cannot expect the location of this taxon in any supertree to be strongly supported.

Thus, we will assume that the sequence of subsets of X satisfies the following *covering property*: For each subset Y of taxa from X of size m (where $m = 3$ for rooted trees or $m = 4$ for unrooted trees), the proportion of subsets X_i that contain Y has strictly positive support as the sequence length (of subsets) increases. More formally, for each such subset Y of X we assume there is some $\epsilon > 0$ and some K sufficiently large for which:

$$(2) \qquad \frac{1}{k}|\{i \le k : Y \subseteq X_i\}| \ge \epsilon \text{ for all } k \ge K.$$

If a subset of taxa, Y, is only found in one or a few trees, and is never seen again in trees that are subsequently added, this property will not hold.

Now, suppose we sample a random tree $\mathcal{T}_i$ on leaf set X_i according to the exponential distribution (1). Let $\mathcal{P}_k = (\mathcal{T}_1, ..., \mathcal{T}_k)$ be the resulting profile of independently sampled trees. The following theorem establishes the statistical consistency of ML supertrees under the exponential model, when the covering condition property holds.

Theorem 3.1. *Given a sequence $X_1, X_2, \ldots$ which satisfies the covering property (2), consider a profile $\mathcal{P}_k = (\mathcal{T}_1, \ldots, \mathcal{T}_k)$, where $\mathcal{T}_i$ is generated independently according to the exponential model (1) with $\beta = \beta_i$ and with generating tree $\mathcal{T}_0$. Suppose that $\beta_i \geq \delta > 0$ for all i. Then the probability that $\mathcal{P}_k$ has a unique ML supertree and that this tree is $\mathcal{T}_0$ tends to 1 as $k \to \infty$.*

Proof. To establish the theorem, using Proposition 8.1 (stated and proved in the Appendix) it is enough to specify for each choice of distinct resolved phylogenetic X–trees $\mathcal{T}_0$ and $\mathcal{T}$, a sequence of events E_k (dependent on $\mathcal{P}_k$) for which, as k grows, E_k has a probability that tends to 1 under the distribution obtained from $\mathcal{T}_0$ and tends to 0 under the distribution obtained from $\mathcal{T}$. Since $\mathcal{T}$ differs from $\mathcal{T}_0$ a subset Y exists of size m ($= 3$ for rooted trees and $= 4$ for unrooted) for which $\mathcal{T}|Y \neq \mathcal{T}_0|Y$. Notice that the covering property (2) implies that

(3) $$\frac{1}{k}|\{i \leq k : \mathcal{T}|X_i \neq \mathcal{T}_0|X_i\}| \geq \epsilon \text{ for all } k \geq K.$$

Let E_k be the event that among all those $i \in \{1, \ldots, k\}$ for which $\mathcal{T}|X_i \neq \mathcal{T}_0|X_i$ we have $\mathcal{T}_i = \mathcal{T}_0|X_i$ more often than $\mathcal{T}_i = \mathcal{T}|X_i$. Now, for a profile generated by $\mathcal{T}_0$ according to the exponential model (1), we have

$$\mathbb{P}_{\mathcal{T}_0}[\mathcal{T}_i = (\mathcal{T}_0|X_i)] = \alpha_i \exp(-\beta_i \cdot 0) = \alpha_i,$$

and for each i for which $\mathcal{T}|X_i \neq \mathcal{T}_0|X_i$, we also have

$$\mathbb{P}_{\mathcal{T}_0}[\mathcal{T}_i = (\mathcal{T}|X_i)] \leq \alpha_i \exp(-\delta d(\mathcal{T}|X_i, \mathcal{T}_0|X_i)).$$

In particular, for each i for which $\mathcal{T}|X_i \neq \mathcal{T}_0|X_i$,

(4) $$\mathbb{P}_{\mathcal{T}_0}[\mathcal{T}_i = (\mathcal{T}_0|X_i)] \geq (1 + \eta)\mathbb{P}_{\mathcal{T}_0}[\mathcal{T}_i = (\mathcal{T}|X_i)] \text{ for some } \eta > 0.$$

Similarly, for a profile generated by $\mathcal{T}$ according to the exponential model (1) and for each i for which $\mathcal{T}|X_i \neq \mathcal{T}_0|X_i$, we have

(5) $$\mathbb{P}_{\mathcal{T}}[\mathcal{T}_i = (\mathcal{T}|X_i)] \geq (1 + \eta)\mathbb{P}_{\mathcal{T}}[\mathcal{T}_i = (\mathcal{T}_0|X_i)] \text{ for some } \eta > 0.$$

By condition (3), there is a positive limiting proportion ($\epsilon > 0$) of i for which $\mathcal{T}|X_i \neq \mathcal{T}_0|X_i$; therefore inequalities (4) and (5) imply that event E_k has a probability tending to 1 (or 0) as $k \to \infty$ for a profile generated by $\mathcal{T}_0$ (or $\mathcal{T}$ respectively) as required. Statistical consistency of ML now follows by Proposition 8.1. $\square$

4. RELATION TO MRP AND ITS STATISTICAL INCONSISTENCY

As shown recently by Bruen and Bryant (2007), there is a close analogy between MRP (Matrix Representation with Parsimony) and consensus tree methods which seek a median tree computed using the SPR (subtree prune and regraft) or TBR (tree bisection and reconnection) metric d (recall that a median tree for a profile $\mathcal{P} = (\mathcal{T}_1, \ldots \mathcal{T}_k)$ of trees that all have same leaf set X, is a tree $\mathcal{T}$ that minimises

the sum $\sum_{i=1}^{k} d(\mathcal{T}, \mathcal{T}_i)$; *cf.* Proposition 2.1). However, the result from Bruen and Bryant (2007) does not guarantee that MRP produces an ML supertree even when $\beta_i = 1$ for all i.

We turn now to the question of the statistical consistency of MRP under the exponential model (1). It can be shown that MRP will be statistically consistent under the covering property (2) in some special cases. Two such cases that can be formally established (details omitted) are: (i) when all the subsets X_i are of size 4, or (ii) when β_i is sufficiently large (in relation to $|X|$). However, in general, we have the following result.

Theorem 4.1. *A $\beta > 0$ exists for which MRP is statistically inconsistent even in the special ('consensus') case where, for all i, X_i is the same set of six taxa and $\beta_i = \beta$. More precisely, for this value of β and with unrooted fully-resolved phylogenetic trees on these (equal) taxon sets, the probability that $\mathcal{T}_0$ is an MRP tree (for a profile of trees generated under (1)) converges to 0 as k tends to infinity.*

Proof. For two unrooted fully-resolved phylogenetic X–trees $\mathcal{T}, \mathcal{T}'$ let $L(\mathcal{T}, \mathcal{T}')$ denote the total parsimony score on $\mathcal{T}'$ of the sequence of splits of $\mathcal{T}$. That is,

$$(6) \qquad L(\mathcal{T}, \mathcal{T}') = \sum_{\sigma \in \Sigma(\mathcal{T})} l(\sigma, \mathcal{T}'),$$

where $\Sigma(\mathcal{T})$ is the set of splits of $\mathcal{T}$ and $l(\sigma, \mathcal{T}')$ is the parsimony score of the split σ on $\mathcal{T}'$ (treating σ as a binary character, (Semple and Steel, 2003)). For any fully-resolved phylogenetic X–tree $\mathcal{T}'$ let $e(\mathcal{T}'; \mathcal{T}_0)$ be the expected total parsimony score on $\mathcal{T}'$ of the sequence of splits of a tree $\mathcal{T}$ randomly generated by $\mathcal{T}_0$ according to the exponential model (1). Then,

$$(7) \qquad e(\mathcal{T}'; \mathcal{T}_0) = \sum_{\mathcal{T}} \alpha \exp(-\beta d(\mathcal{T}, \mathcal{T}_0)) \cdot L(\mathcal{T}, \mathcal{T}').$$

To establish Theorem 4.1, it is enough to show, for some $\beta > 0$ and for two unrooted fully-resolved trees $\mathcal{T}_0, \mathcal{T}_1$ on $X = \{1, \ldots, 6\}$, that $e(\mathcal{T}_0; \mathcal{T}_0) - e(\mathcal{T}_1; \mathcal{T}_0) > 0$, since if $\mathcal{T}_0$ is the generating tree, then $\mathcal{T}_1$ will be favored over $\mathcal{T}_0$ by MRP. We first show that this can occur when $\beta = 0$. In that case, $\alpha \exp(-\beta d(\mathcal{T}, \mathcal{T}_0)) = 1/105$ for all $\mathcal{T}$ (there are 105 unrooted fully-resolved phylogenetic trees on X) and so, by (6) and (7),

$$e(\mathcal{T}_0; \mathcal{T}_0) - e(\mathcal{T}_1; \mathcal{T}_0) = \frac{1}{105} \sum_{\mathcal{T}} (L(\mathcal{T}, \mathcal{T}_0) - L(\mathcal{T}, \mathcal{T}_1)) = \frac{1}{105} \sum_{\sigma} n(\sigma) \cdot (l(\sigma, \mathcal{T}_0) - l(\sigma, \mathcal{T}_1)),$$

where $n(\sigma)$ is the number of unrooted fully-resolved phylogenetic X–trees containing split σ and the summation is over all the splits of $X = \{1, \ldots, 6\}$. Now suppose $\mathcal{T}_0$ has a symmetric shape (i.e. an unrooted fully-resolved tree of six leaves with three cherries) and $\mathcal{T}_1$ has a pectinate shape (i.e. an unrooted fully-resolved tree of six leaves with two cherries). Then, by using earlier results (Steel et al., 1992, Table 3) and basic counting arguments, it can be shown that

$$e(\mathcal{T}_0; \mathcal{T}_0) - e(\mathcal{T}_1; \mathcal{T}_0) = \frac{1}{21}.$$

So far, we have assumed that $\beta = 0$, however, $e(\mathcal{T}_0; \mathcal{T}_0) - e(\mathcal{T}_1; \mathcal{T}_0)$ is a continuous function of β so a strictly positive value of β exists for which

$$e(\mathcal{T}_0; \mathcal{T}_0) - e(\mathcal{T}_1; \mathcal{T}_0) \geq \frac{1}{10}.$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ $\square$

An interesting theoretical question is whether a value $s \in (0, 1)$ exists for which MRP is statistically consistent (for arbitrarily large taxon sets) under the conditions of Theorem 4.1, whenever $\beta \geq s$.

5. Technical remarks

Extension to trees with polytomies.

We can easily modify the ML process if some of the input trees are not fully-resolved. For a general phylogenetic tree t_i (possibly with polytomies) on taxon set $X_i \subseteq X$, and a generating fully-resolved phylogenetic tree $\mathcal{T}$ on taxon set X, let $\phi(t_i|\mathcal{T})$ be the probability of the event that the tree $\mathcal{T}_i$ that $\mathcal{T}$ generates under the exponential model is a refinement of t_i. More precisely,

$$\phi(t_i|\mathcal{T}) = \sum_{\mathcal{T}_i \geq t_i} \alpha_i \exp(-\beta_i d(\mathcal{T}_i, \mathcal{T}|X'))$$

where $\mathcal{T}_i \geq t_i$ indicates that the (fully-resolved) tree $\mathcal{T}_i$ contains all the splits present in t_i, and has the same leaf set (X_i). Notice that $\phi(t_i|\mathcal{T})$ is not a probability distribution on phylogenetic trees with the leaf set X' (its sum is > 1). Nevertheless, given a profile $\mathcal{P} = (t_1, \ldots, t_k)$ of phylogenetic trees (some or all of which may have polytomies), we can perform ML to select the tree $\mathcal{T}$ that maximises the joint probability $\prod_{i=1}^{k} \phi(t_i|\mathcal{T})$ of the events $\mathcal{T}_i \geq t_i$ for $i = 1, \ldots, k$.

An alternative perspective on ML supertrees for certain tree metrics.

We point out an alternative way of viewing this ML procedure applied to a profile $\mathcal{P} = (\mathcal{T}_1, \ldots, \mathcal{T}_k)$ when d is one of two well-known metrics on trees (SPR and TBR). Suppose that we were to extend each tree $\mathcal{T}_i$ in $\mathcal{P}$ to a tree $\mathcal{T}_i'$ on the full set of taxa (X). We could regard the placement of those taxa that are 'missing' in $\mathcal{T}_i$ (namely the taxa in $X - X_i$) to form a tree $\mathcal{T}_i'$ on the full leaf set X to be 'nuisance parameters' in a maximum likelihood framework (under the exponential model), and thereby seek to find the tree $\mathcal{T}$ and extensions $(\mathcal{T}_1', \ldots, \mathcal{T}_k')$ to maximise the joint probability:

$$\mathbb{P}_{\mathcal{T}}[(\mathcal{T}_1', \mathcal{T}_2', \ldots, \mathcal{T}_k')] \text{ subject to } \mathcal{T}_i = \mathcal{T}_i'|X_i \text{ for all } i.$$

We call such a tree $\mathcal{T}$ an *extended ML tree* for the profile $\mathcal{P}$.

Proposition 5.1. *For $d = SPR$ or $d = TBR$, and any profile $\mathcal{P}$ of fully-resolved, unrooted phylogenetic trees, the extended ML tree(s) for $\mathcal{P}$ coincides precisely with the ML tree(s) for $\mathcal{P}$.*

Proof. For $d = SPR$ or $d = TBR$, it can be shown that for any resolved unrooted phylogenetic trees $\mathcal{T}$ on leaf set X, and $\mathcal{T}_i$ on leaf set X_i, that:

$$(8) \qquad \min\{d(\mathcal{T}_i', \mathcal{T}) : \mathcal{T}_i'|X_i = \mathcal{T}_i\} = d(\mathcal{T}_i, \mathcal{T}|X_i).$$

The result now follows by Proposition 2.1   □

Note that Equation 8 does not necessarily hold for other tree metrics (such as the NNI (nearest-neighbor interchange) or the partition (Robinson-Foulds) metric).

6. Statistical consistency of ML species supertrees from multiple gene trees

A current problem in phylogenetics is how best to infer species trees from gene trees (Degnan and Rosenberg, 2006; Gadagkar et al., 2005; Liu and Pearl, 2007). Even in the consensus setting (i.e. when the set of taxa for each gene tree is the complete set of taxa under study), Degnan and Rosenberg (2006) have demonstrated how incomplete lineage sorting on gene trees can mean that the most likely topology for a gene tree can differ from the underlying species tree (for any certain rooted phylogenetic trees on four taxa and for all rooted phylogenetic trees on five or more taxa). This surprising result implies that simplistic 'majority rule' approaches to finding a consensus species tree can be problematic.

The phenomenon described by Degnan and Rosenberg (2006) is based on the coalescent model for studying lineage sorting in evolving populations. The surprising behavior arises only when the effective population sizes and the branch lengths of the species tree are in appropriate ranges. Moreover, for 3–taxon trees, the most probable gene tree topology always agrees with the species tree topology. Nevertheless, the fact that larger gene trees can favour an incorrect species tree might easily complicate some standard statistical approaches.

In this section, we show how, despite the phenomena described above (from Degnan and Rosenberg, 2006), and even in the more general supertree setting (where some gene trees may have some missing taxa), a maximum likelihood approach to supertree construction of a species tree from gene trees is statistically consistent. Moreover, we frame this approach so that it is sufficiently general to also allow for error in the reconstruction of the gene trees (as arises under the exponential model).

Consider a model M that has a generating tree topology $\mathcal{T}$ as its sole underlying parameter. Such a model will typically derive from a more complex model containing other parameters (such as branch lengths, population sizes and so forth), but we will assume that these have a prior distribution and that they have been integrated out, so our model has just one parameter - the tree topology. We say that M satisfies the property of *basic centrality* if, for all subsets of Y of X of size m ($= 3$ for rooted trees and $= 4$ for unrooted trees), we have:

$$(9) \qquad \mathbb{P}_{\mathcal{T}}[\mathcal{T}|Y] \geq (1 + \eta)\mathbb{P}_{\mathcal{T}}[\mathcal{T}']$$

for all trees $\mathcal{T}'$ on leaf set Y that are different from $\mathcal{T}|Y$, and where $\eta > 0$. For example, the exponential model (1) satisfies basic centrality, since (9) holds for *all* subsets Y of X. For lineage sorting (with a prior distribution on ancestral population sizes and branch lengths), the property holds, but only because (9) holds for the subsets Y of X of size 3, as we describe shortly. Firstly, however, we state a statistical consistency result that extends Theorem 3.1.

Proposition 6.1. *Given a sequence $X_1, X_2, \ldots$ which satisfies the covering property (2), consider a profile $\mathcal{P}_k = (\mathcal{T}_1, \ldots, \mathcal{T}_k)$, where $\mathcal{T}_i$ is generated independently according to a model that satisfies the basic centrality property (with $\mathcal{T} = \mathcal{T}_i$, and $\eta_i > \delta > 0$ for all i) and with generating tree $\mathcal{T}_0$. Then the probability that $\mathcal{P}_k$ has a unique ML supertree and that this tree is $\mathcal{T}_0$ tends to 1 as $k \to \infty$.*

Proof. The proof is similar to the proof of Theorem 3.1, the essential difference being a modification to the way the events E_k are defined. Given $\mathcal{T}_0$ and $\mathcal{T}$ (as in the proof of Theorem 3.1), let E_k be the event that for each $i \in \{1, \ldots, k\}$ for which $Y \subseteq X_i$ we have $\mathcal{T}_i|Y = \mathcal{T}_0|Y$ more often than $\mathcal{T}_i|Y = \mathcal{T}|Y$. Then (as in the proof of Theorem 3.1) as k grows, E_k has a probability that converges to 1 if $\mathcal{T}_0$ is the generating tree, and a probability that converges to 0 if $\mathcal{T}$ is the generating tree. The theorem now follows by Proposition 8.1. $\square$

We now apply this result in a supertree setting where we have the compounding effect of *two* sources of error: (i) error in using the true gene tree to represent the true species tree, due to lineage sorting, and (ii) error in reconstructing the correct gene tree, modeled by the exponential model (1). We claim that an ML approach to inferring a species tree in the presence of these two sources of error is still statistically consistent (under the coalescent and exponential model (1), and assuming the covering property), due to the following argument, which justifies the basic centrality property.

Consider a rooted fully-resolved species tree $\mathcal{T}$ on X and a rooted full-resolved gene tree $\mathcal{T}'$ on Y, where Y is a subset of X of size 3 (note that we are here identifying the taxa in the gene tree with taxa in the species tree). Then the probability of observing $\mathcal{T}'$ under the combination of these two sources of error (treated independently) from a generating species tree $\mathcal{T}$ can be written as

$$\mathbb{P}_{\mathcal{T}}[\mathcal{T}'] = \sum_{\mathcal{T}''} \mathbb{P}_{\mathcal{T}}^c[\mathcal{T}''] \mathbb{P}_{\mathcal{T}''}[\mathcal{T}'],$$

where the summation is over the three rooted fully-resolved gene trees on taxon set Y, $\mathbb{P}_{\mathcal{T}}^c[\mathcal{T}'']$ is the probability that species tree $\mathcal{T}$ gives rise to gene tree $\mathcal{T}''$ for the taxa on Y (under lineage sorting according to the coalescent model), and $\mathbb{P}_{\mathcal{T}''}[\mathcal{T}']$ is the probability that a generating gene tree $\mathcal{T}''$ produces $\mathcal{T}'$, as given by the exponential model (1). Now, considering lineage sorting under the coalescent model, we have $\mathbb{P}_{\mathcal{T}}^c[\mathcal{T}|Y] = \frac{1}{3}(1 + 2\tau)$ for $\tau \in (0, 1)$ while $\mathbb{P}_{\mathcal{T}}^c[\mathcal{T}''] = \frac{1}{3}(1 - \tau)$ for the two other choices of $\mathcal{T}'' \neq \mathcal{T}|Y$ (see e.g., Rosenberg, 2002; Tajima, 1983). Furthermore, under the exponential model (1), and assuming, without loss of generality, that d takes the value 0 or 1 for each pair of 3-taxon trees, we have $\mathbb{P}_{\mathcal{T}''}[\mathcal{T}''] = \alpha$, while $\mathbb{P}_{\mathcal{T}''}[\mathcal{T}'] = \alpha e^{-\beta}$ for the other two choices of $\mathcal{T}' \neq \mathcal{T}''$ (and $\alpha = (1 + 2e^{-\beta})^{-1}$).

Combining these relationships, we obtain:

$$\mathbb{P}_{\mathcal{T}}[\mathcal{T}|Y] = \frac{\alpha}{3}\left(1 + 2\tau + 2(1-\tau)e^{-\beta}\right),$$

while for the other two choices of $\mathcal{T}' \neq \mathcal{T}|Y$, we have

$$\mathbb{P}_{\mathcal{T}}[\mathcal{T}'] = \frac{\alpha}{3}\left((1 + 2\tau)e^{-\beta} + (1-\tau)(1 + e^{-\beta})\right).$$

For any given $\beta, \tau > 0$, these last two equations imply that for some $\eta > 0$, and for the two choices of $\mathcal{T}' \neq \mathcal{T}|Y$, we have

$$\mathbb{P}_{\mathcal{T}}[\mathcal{T}|Y] \geq (1 + \eta)\mathbb{P}_{\mathcal{T}}[\mathcal{T}']$$

(in fact, routine algebra shows that we can take $\eta = 3\tau(1 - e^{-\beta})$). Taking the value of η that is minimal for all subsets Y of X of size 3 establishes the basic centrality property in this setting.

7. Discussion

To develop a likelihood-based supertree reconstruction method, it is necessary to define a model that delivers the probability of obtaining a series of subtree topologies, given a hypothesised supertree. We have chosen a very simple yet intuitive probability function whereby the probability of observing a 'wrong' subtree (i.e. one where the topology differs from that of a pruned supertree) decreases exponentially as its topology becomes increasingly distant from that of the hypothesised supertree. Consequently, the ML supertree can be estimated even when the constituent subtrees have conflicting topological signals.

Our approach is model-based, but one may reasonably ask whether the model described here is a biologically realistic one. We suggest that it is. For one thing, we expect, for a variety of reasons, to see conflicts between the topologies of subtrees and the reconstructed supertree. With gene sequences obtained from different species, for instance, incomplete lineage sorting and ancestral heterozygosity frequently lead to differences between gene trees and species trees. Convergent and parallel evolution can confound phylogenetic reconstruction, as can long-branch attraction. We have chosen to use the exponential distribution to describe this steady decrease in probabilities as the distances between subtrees and supertrees increase. The value of using the exponential distribution lies in the ease with which it can be manipulated when we compute log-likelihoods. However, we suggest that one fruitful research project may be to explore other possible probability distributions and, for that matter, other tree-to-tree distance metrics.

The likelihood framework provides an additional benefit: a rich body of statistical and phylogenetic methods already use likelihood. Moreover, statistical consistency holds for maximum likelihood supertrees under weak conditions, in contrast to MRP, which can be inconsistent in some cases. We also show that the ML supertree approach developed here provides a statistically consistent strategy for combining gene trees even when there is the possibility that these trees may be different from the true species tree. An obvious application of ML supertrees will be their use in statistical tests of topological hypotheses, and we already know how to do this with standard ML phylogenies (Goldman et al., 2000).

We also recognise that our particular likelihood implementation is closely related to the 'Majority-Rule(-) Supertree' construction proposed by Cotton and Wilkinson (2007)). More precisely, when the tree metric is the symmetric difference (Robinson-Foulds) metric, then the Marjority-Rule(-) Supertree is, in effect, the strict consensus of our ML supertrees. However, the approach in Cotton and Wilkinson (2007) is quite different: they show how to extend majority-rule from the consensus to the supertree setting. Nonetheless, they converge on the same optimality criterion that we use, i.e. a supertree that minimises the sum of distances to a set of trees. One should not be surprised that the same optimality criterion can emerge from different conceptual bases. With standard phylogenetic reconstruction, choosing the tree that minimises the number of evolutionary changes can be justified philosophically (with the principle of maximum parsimony) as a consensus method (Bruen and Bryant, 2007), or by using an explicitly statistical approach (e.g. likelihood (Steel and Penny, 2000)).

We have not discussed algorithms to search for ML supertrees. Instead, we direct readers to the discussion in Cotton and Wilkinson (2007), since the criterion we use is similar to theirs.

8. Appendix: Consistency of ML for general (non-i.i.d.) sequences

Here we describe a convenient way to establish the statistical consistency of maximum likelihood when we have a sequence of observations that may not be independent or identically distributed. We frame this discussion generally, as the result may be useful for other problems. In particular, in this result, we do not need to assume the sequence samples are independent (though in our applications, they are), nor identically distributed (in our applications, they are not). Suppose we have a sequence of random variables $Y_1, Y_2, \ldots$ that are generated by some process that depends on an underlying discrete parameter a that can take values in some finite set A. In our setting, the Y_i's are trees constructed from different data sets (e.g. gene trees), while a is the generating species tree topology. We assume that the model specifies the probability distribution of $(Y_1, \ldots, Y_k)$ given (just) a – for example, in our tree setting this would mean specifying prior distributions on the branch lengths and other parameters of interest (eg. ancestral population sizes) and integrating with respect to these priors.

Given an actual sequence $(y_1, ..., y_k)$ of observations, the *maximimum likelihood (ML) estimate* of the discrete parameter is the value a that maximises the joint probability

$$\mathbb{P}_a[Y_1 = y_1, \ldots, Y_k = y_k]$$

(i.e. the probability that the process with parameter a generates $(y_1, \ldots, y_k)$). Now suppose that the sequence $Y_1, \ldots Y_k, \ldots$ is generated by a_0. We would like the probability that the ML estimate is equal to a_0 to converge to 1 as k increases. If this holds for all choice of $a_0 \in A$, then ML is *statistically consistent*. The following result provides an convenient way to establish this; indeed, it characterises the statistical consistency of ML.

Proposition 8.1. *In the general set-up described above, ML is statistically consistent if and only if the following condition holds: for any two distinct elements $a, b \in A$, we can construct a sequence of events $E_1, E_2, \ldots$, where E_k is dependent on $(Y_1, \ldots Y_k)$, for which, as $k \to \infty$:*

   (i) *the probability of E_k under the model with parameter a converges to 1.*
   (ii) *the probability of E_k under the model with parameter b converges to 0.*

Proof. The 'only if' direction is easy: Suppose ML is statistically consistent and $a, b \in A$ are distinct. Let E_k be the event that a is the unique maximum likelihood estimate obtained from $(Y_1, \ldots, Y_k)$. Then E_k satisfies conditions (i) and (ii).

For the converse direction, recall that the *variation distance* between two probability distributions p, q on a finite set W is

$$\max_{E \subset W} |\mathbb{P}_p(E) - \mathbb{P}_q(E)|$$

where $\mathbb{P}_p(E) = \sum_{w \in E} p(w)$ is the probability of event E under distribution p (similarly for $\mathbb{P}_q(E)$). This variation distance can also be written as $\frac{1}{2}\|p - q\|_1$, where $\|p - q\|_1 = \sum_{w \in W} |p(w) - q(w)|$ is the l_1 distance between p and q. Thus, if we let $d^{(k)}(a, b)$ denote the l_1 distance between the probability distribution on $(Y_1, \ldots, Y_k)$ induced by a and by b, then conditions (i) and (ii) imply that

$$(10) \qquad\qquad \lim_{k \to \infty} \frac{1}{2} d^{(k)}(a, b) = 1.$$

Now, by the first part (Eqn. 3.1) of Theorem 3.2 of (Steel and Székely, 2002), the probability that the ML estimate is the value of A that generates the sequence $(Y_1, \ldots, Y_k)$ is at least $1 - \sum_{b \neq a}(1 - \frac{1}{2}d^{(k)}(a, b))$ and so, by (10), this probability converges to 1 as $k \to \infty$. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

References

Baum, B. R. and M. A. Ragan. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10.

Bininda-Emonds, O. R. P., M. Cardillo, K. E. Jones, R. D. E. MacPhee, R. M. D. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman, and A. Purvis. 2007. The delayed rise of present-day mammals. Nature 446:507–512.

Bruen, T. and D. Bryant. 2007. Parsimony as consensus. Syst. Biol. (in press).

Cotton, J. A. and R. D. M. Page. 2004. Tangled tales from multiple markers. Chapter 5 *in* Phylogenetic Supertrees (O. R. P. Bininda-Emonds, ed.). Kluwer Academic Publishers, Dordrecht, The Netherlands.

Cotton, J. A. and M. Wilkinson. 2007. Majority-rule supertrees. Syst. Biol. 56:445–452.

Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2006 May 2:e68.

Gadagkar, S. R., M. S. Rosenberg, and S. Kumar. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J. Exper. Zool. 304B:64–74.

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.

Liu, L. and D. K. Pearl. 2007. Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Rodrigo, A. G. 1996. On combining cladograms. Taxon 45:267–274.

Ronquist, F., J. J. Huelsenbeck, and T. Britton. 2004. Bayesian supertrees. Chapter 9 *in* Phylogenetic Supertrees (O. R. P. Bininda-Emonds, ed.). Kluwer Academic Publishers, Dordrecht, The Netherlands.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Pop. Biol. 61:225–247.

Ross, H. A. and A. G. Rodrigo. 2004. An assessment of Matrix Representation with compatibility in supertree reconstruction. Chapter 2 *in* Phylogenetic Supertrees (O. R. P. Bininda-Emonds, ed.). Kluwer Academic Publishers, Dordrecht, The Netherlands.

Semple, C. and M. Steel. 2003. Phylogenetics. Oxford University Press.

Steel, M. and D. Penny. 2000. Parsimony, likelihood and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

Steel, M. and L. A. Székely. 2002. Inverting random functions (ii): explicit bounds for discrete maximum likelihood estimation, with applications. SIAM J. Discr. Math. 15:562–575.

Steel, M. A., M. D. Hendy, and D. Penny. 1992. Significance of the length of the shortest tree. J. Classif. 9:71–90.

Tajima, F. 1983. Evolutionary relationships of dna sequences in finite populations. Genetics 105:437–460.

MS: Allan Wilson Centre for Molecular Ecology and Evolution, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: `m.steel@math.canterbury.ac.nz`

AR: Bioinformatics Institute, University of Auckland, New Zealand, and Laboratoire d'Informatique, de Robotique et de Microelectronique de Montpellier, France.

E-mail address: `a.rodrigo@auckland.ac.nz`