Species Tree and Most Likely Gene Tree

MATH285K - Spring 2010 *Presenter: Gabriela Cybis*

Sorting of gene lineages at speciation can lead to disagreement between the species tree and the gene tree. But it is usually assumed that the most likely gene tree for a given set of species coincides with the species tree. This is what motivates using gene information to estimate the species tree. This paper shows that, under the coalescent model of within species evolution, for many trees there exists a set of branch lengths for which the most likely gene tree to evolve along the branches of a species tree differs from the species topology. Thus, accumulating information from many gene trees can lead to convergence of the estimate to the wrong tree (as the number of genes increases).

Consider the definition:

Definition 1:

(i)A gene tree topology is *anomalous* for a species tree $\sigma = (\psi, \lambda)$ if $P_\sigma(G = g) > P_\sigma(G = \psi)$.

(ii) A topology ψ *produces anomalies* if there exists a vector of branch lengths λ such that the species tree $\sigma = (\psi, \lambda)$ has at least one anomalous gene tree.

(iii) The *anomaly zone* for a topology ψ is the set of vectors of branch lengths λ for which $\sigma = (\psi, \lambda)$ has at least one anomalous gene tree.

Anomalous trees cannot happen with 3 taxa:
Denoting the length of the internal branch by λ, the probability that the gene tree has the same topology as the species tree is $1 - (2/3)e^{-\lambda}$. This value always exceeds the probability that the gene topology matches one of the other two topologies $(1/3)e^{-\lambda}$. [2]

With four taxa, asymmetric trees with short internal branches can produce anomalies:
Make the internal branches short enough that with high probability all coalescent events happen above the tree root. In this case, all symmetric topologies have probability 1/9, and all asymmetric topologies have probability 1/18. Thus, asymmetric topologies can produce anomalies. [2]

Proposition 2: Any species tree topology with $n \geq 5$ taxa produces anomalies.

To prove the proposition, we need the following definition and lemmas:

Definition 3: A labeled topology L_n for n taxa is *n-maximally probable* if its probability under the Yule model of random branching is greater or equal to that of any other labeled topology for n taxa.

In the Yule model, each lineage has the same probability of dividing, giving rise to two lineages. It is the equivalent to the coalescent model, but inversed in time [3].

Lemma 4: For $n \geq 4$, any species tree that is not n-maximally probable produces anomalies.

Lemma 4 is proved by making the internal branches of the trees short enough, that with hight probability all gene lines have not coalesced before the tree root. Then the most likely gene tree is the n-maximally probable tree.

Lemma 5: For $n \in \{5, 6, 7, 8\}$, any species tree topology that is n-maximally probable produces anomalies.

Lemma 5 is proved by exhaustively looking at all n-maximally probable topologies for $n \in \{5, 6, 7, 8\}$, and showing that they can all produce anomalies.

Idea of the proof of Proposition 3: We want to show that for $n \geq 9$ any n-maximally probable tree topology can produce anomalies. We procede by induction:
Any n-maximally probable species tree can be divided at the root in two sub-trees, each with smaller number of species. But we know that at least one of these trees can produce anomalies (since for one of them $n \geq 5$). If we choose branch lengths accordingly, then one of the subtrees is likely to produce anomalies, and the other is not. With these branch lengths, the species tree topology has an anomalous gene tree. □

Results of this paper predict situations where there existis a set of labeled topologies where the most likely gene tree for one topology is the other and vice versa. The authors call these sets of trees *wicked forests*. In this case, the more data you gather from one gene tree, the larger your evidence for the other.

The work in the paper assumes a perfect method for inferring the gene tree. But gene trees are usually estimated from mutational data. The anomaly zones for different topologies usually include short internal branches. But if these branches

are short, frequently there isn't enough information about that specific branching because very few mutations happen along it. In those cases, anomalous gene trees aren't even an issue.

So in actual sequence analysis, anomalous gene trees may only come up when these short internal branches have high mutation rates. Simulation studies could be useful to evaluate how often the problem of anomalous gene trees arises.

Solutions to the anomalous gene tree problem might include:

- Sampling many individuals from each species. This would increase the probability of coalescent events in short branches, one of the main drivers of anomalies. The solution would only be good when external branches aren't too long, otherwise all coalescent events happen before the internal branches that need to be resolved.

- Since 3 species trees cannot create anomalies, one could resolve first all 3 species relationships, and then merge this information to obtain the species tree.

References

[1] Degnan J, Rosenberg, N (2006) Discordance of Species Trees with Their Most Likely Gene Tree. PLOS Genetics 05(2): e68

[2] Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5: 568 - 583.

[3] Steel M, McKenzie A (2001) Properties of phylogenetic trees generated by Yule-type speciation models. Math Biosci 170: 91 - 112